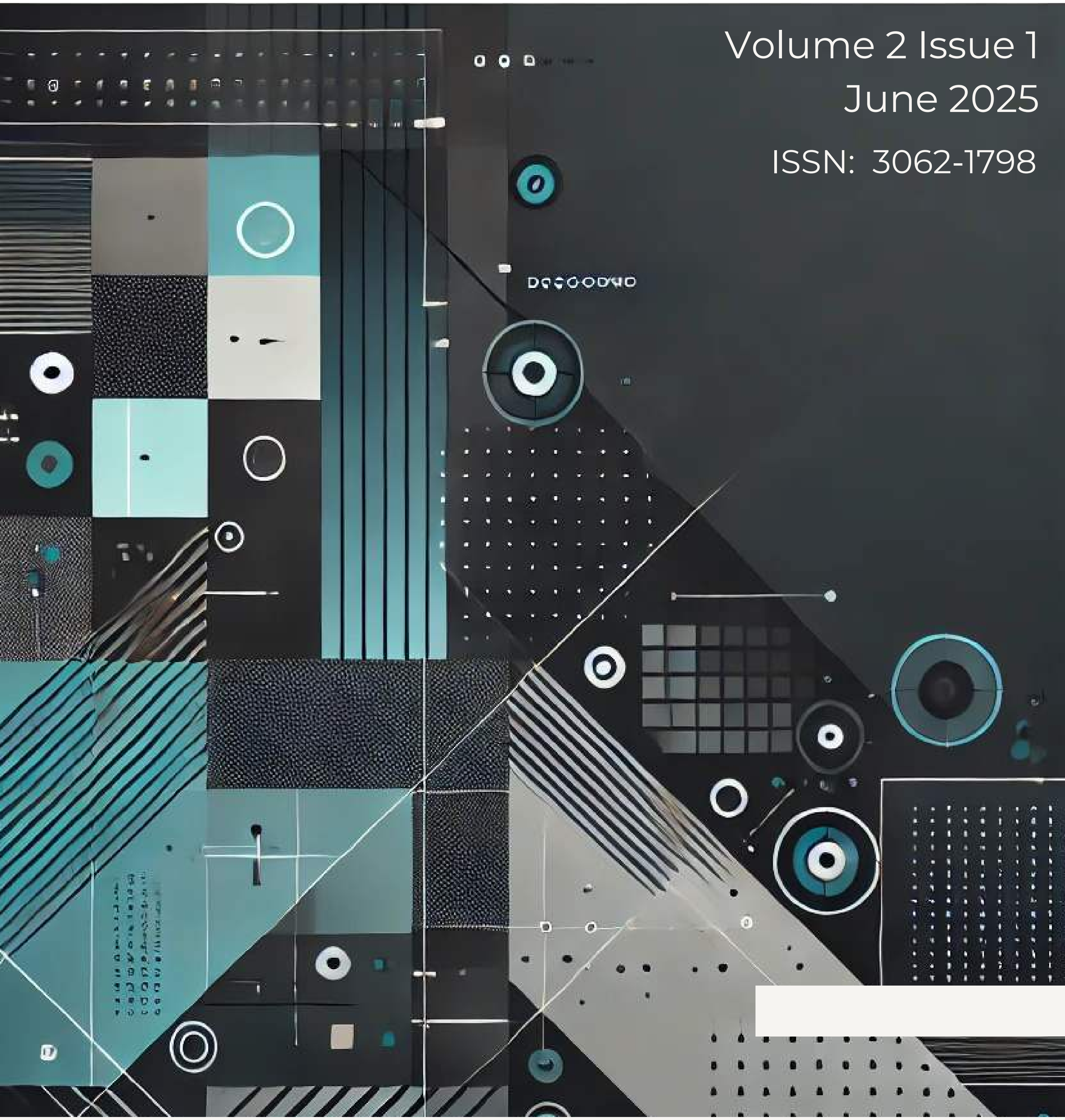
Research & Design



Volume 2 Issue 1
June 2025
ISSN: 3062-1798

MIM Research & Publishing

Research & Design (REDE) is a multidisciplinary, peer-reviewed, open-access journal committed to bridging the gap between research and design. Inspired by the fundamental problem-solving mechanisms of science, REDE fosters innovation by publishing studies that connect theoretical exploration with practical applications. Our mission is to advance knowledge and provide a platform where research-driven solutions meet real-world challenges.

Supported by the MIM Research Group, **REDE offers a free and open-access publication model**, ensuring that all published articles are freely available to the global research community. Authors are not required to pay any fees, reflecting the journal's dedication to inclusivity and equitable knowledge dissemination.

The journal welcomes contributions from diverse disciplines, emphasizing the interconnectedness of research and design across various fields. Original research articles, reviews, technical notes, and short communications that address complex societal issues are encouraged. Each submission undergoes a rigorous peer-review process to ensure academic excellence and impact.

REDE covers a wide array of fields, reflecting its commitment to multidisciplinary collaboration. From traditional engineering disciplines to contemporary topics like gamification and data science, the journal provides a home for innovative studies that address modern challenges.

We invite you to explore our diverse content and contribute to advancing research and design in your field.

The topics covered in REDE include:

- Aesthetics
- Built Environment
- Economics and Finance
- Social Sciences
- Engineering Sciences
- Data Science
- Game & Gamification
- Industrial Design
- Interface Design

Abstracting and Indexing
Please visit http://www.jrede.org/ for more information.

The journal of

RESEARCH & DESIGN

Research Group

Research Group
Research & Design
www.jrede.org

Research Group
Research & Design
www.jrede.org

Editorial Board Members

2025 Advisory Board Members (in progress)

Research & Design
www.jrede.org

In This Issue

Free access to tables of content, abstracts and full text of papers for web visitors.

Research & Design
MIM Research Group Publications

http://www.jrede.org/

Research & Design
www.jrede.org

Research Article

Cultural dimensions in the perception of success: Comparative analysis of word associations across languages using LLM word embedding

Hayri Baytan Ozmen*[a]

Faculty of Engineering and Natural Sciences, Uşak University, Uşak, Türkiye.

Article Info

Article History:

Received:
14 Apr 2025

Accepted:
26 May 2025

Keywords:

Ambition;
computational social science;
cross-cultural analysis;
cultural values;
effort;
semantic similarity

Abstract

This study investigates how the concept of success is semantically associated with culturally salient attributes across nine languages using monolingual embeddings derived from large language models. Twelve key attributes—effort, ambition, talent, collaboration, happiness, luck, creativity, discipline, education, status, wealth, and respect—were analyzed based on their cosine distances to "success." Lower distances indicate stronger conceptual ties. Findings reveal that effort is universally central, appearing closest to success in all languages studied. Cultural nuances were evident: collaboration and ambition showed varying levels of association, with collectivist cultures such as Chinese and Arabic emphasizing collaboration, while individualistic cultures like Finnish and English highlighted ambition. Talent and happiness emerged as significant in specific contexts, particularly in Finnish, Russian, and Turkish corpora. Luck showed a stronger link in European languages like German and Russian, suggesting higher attribution of success to chance. Conversely, external markers like status and wealth showed weaker associations overall. These results offer a data-driven, cross-linguistic perspective on how success is framed within different cultural value systems.

1. Introduction

Language extends far beyond its role as a mere medium of communication; it serves as a profound reflection of the cultural values and social paradigms intrinsic to the communities that speak it. Each language embodies the shared history, beliefs, and practices of a culture, thus acting as a repository of collective identity and a framework through which individuals understand their world. The interconnection between language and culture suggests that the words we choose to convey ideas are heavily laden with cultural significance, shaping and mirroring societal norms and values.

The significance of this relationship is supported by traditional sociological frameworks, such as Hofstede's dimensions of culture and the Inglehart–Welzel cultural map [1, 2]. Hofstede's work elucidates how different cultures prioritize values differently, distinguishing between individualism and collectivism, uncertainty avoidance, power distance, masculinity versus femininity, long-term versus short-term orientation, and indulgence versus restraint [3]. These dimensions highlight that language is a vessel that

Corresponding author: baytan.ozmen@usak.edu.tr
[a]orcid.org/0000-0001-6750-8632
DOI: http://dx.doi.org/10.17515/rede2025-001so0414rs

carries the weight of cultural priorities, influencing how foundational concepts such as “success” are perceived and articulated across different societies.

Similarly, Inglehart and Welzel’s cultural map positions countries along various axes such as traditional versus secular-rational values and survival versus self-expression values [2]. This mapping enables scholars to uncover deeper patterns of cultural differentiation and resonance, providing a framework for analyzing how language encapsulates these broader existential orientations. As such, cultural values become manifest in linguistic expressions and word choices, highlighting the intersections between linguistic constructs and sociocultural frameworks.

The sociological discourse surrounding culture and language emphasizes the importance of examining these constructs to gain nuanced insights into human behavior. McGuigan and Moran assert that cultural materialism facilitates a deeper understanding of social phenomena by linking cultural expressions to the specific socio-economic conditions and power dynamics from which they arise [4]. This reinforces the need for a holistic approach in understanding language as a cultural artifact; it is not only shaped by culture but also actively shapes cultural perceptions and behaviors.

Furthermore, the exploration of how language reflects cultural values enhances our comprehension of cultural autonomy and the myriad ways in which social structures influence linguistic practices. Watts highlights how sociological investigations into culture demonstrate its dual role as both an independent variable, influencing social structures, and a dependent variable, shaped by prevailing social contexts [5]. Hence, analyzing word meanings and spaces in word meanings across languages provides an empirical basis for exploring these intricate cultural dynamics.

1.1 Problem Statement and Research Question

The relationship between language and culture has long been recognized as a critical area of study, illuminating how linguistic constructs shape and reflect cultural values and social norms. Focusing specifically on the concept of "success," this research aims to quantitatively compare the semantic associations of success with relevant cultural concepts across distinct languages using word embeddings. Traditional frameworks for understanding culture, such as Hofstede’s cultural dimensions, provide foundational insights into how various dimensions, including individualism versus collectivism and uncertainty avoidance, can influence perceptions of success across different societies [6]. However, some scholars critique these dimensions for their limitations, arguing that they do not account for the complexities of language differences, religious influences, and political contexts which further nuance the cultural landscape [7, 8].

Understanding success through the lens of word embeddings allows for an empirical analysis of closely related concepts in different languages, thus transcending the boundaries of traditional cultural metrics. Previous studies have shown that applying Hofstede's dimensions in exploring cultural disparities in contextual perceptions can yield significant insights into behaviors and decision-making processes across cultures [9]. Considering the significant role of language in shaping cognitive frameworks, employing word embeddings can reveal underlying cultural associations and the linguistic constructs of success that may vary from one culture to another [7].

Word embeddings are a foundational tool in natural language processing (NLP) that represent words as dense vectors in a continuous multidimensional space, capturing semantic and syntactic relationships based on contextual similarity [10]. These embeddings are typically learned from large text corpora using neural network-based models and are encoded in matrices where each row corresponds to a word and its position in the embedding space. In the context of social science research, such as cross-cultural studies of meaning, word embeddings allow for the quantitative comparison of concepts across languages by measuring the similarity between vectors [11]. This method facilitates the analysis of how abstract social constructs—like "success"—are conceptualized differently across cultures by examining their proximity to culturally loaded terms in high-dimensional space [12]. Importantly, embedding matrices provide a non-invasive, data-driven approach to mapping collective semantic associations, making them particularly suitable for exploring cultural dimensions in psychological and sociolinguistic research.

By quantitatively assessing the distances between the word "success" and its culturally relevant counterparts, this study aims to deepen the understanding of how cultural contexts influence conceptual interpretations. Such analysis is vital not only for theoretical frameworks of cultural psychology and linguistic anthropology but also for practical applications in cross-cultural communication and international business where understanding diverse cultural perspectives can influence effectiveness and receptivity [13, 14]. Therefore, this research seeks to bridge the gap in existing literature by providing a quantitative foundation for exploring cultural differences in the conceptualization of success through word embeddings.

The investigation into the cultural dimensions of the concept of "success" entails understanding various attributes that are strongly associated with success in different cultural contexts. Each culture may prioritize distinct characteristics in their definitions of success, highlighting the interplay between cultural and linguistic frameworks. Recognizing how these attributes manifest elucidates the diversity in societal values and aspirations, facilitating a deeper comprehension of what qualities individuals in different cultures consider essential for achieving success.

Additionally, exploring how these associations differ across cultures unveils patterns of cultural divergence and convergence. By examining the embedding distances among word representations in different languages, the research will reveal specific linguistic nuances that illustrate varying cultural interpretations. This comparative analysis will not only illuminate the richness of cultural perspectives on success but also contribute to a broader understanding of how language captures these complexities. Through the lens of cross-cultural psychology, this inquiry aligns with the growing recognition that language acts as a significant cue for cultural mindsets, revealing underlying values and beliefs [15]. The insights gained from such an analysis will underline the intricate relationship between language and cultural identity, significantly enriching discussions on the sociology of culture and the psychological dimensions of human behavior [16].

In short, this research endeavors to uncover both the attributes associated with success in various cultures and the variances in these associations, thereby fostering a more nuanced understanding of how cultural values are reflected in language and how they influence perceptions of key concepts such as success.

1.2 Overview of Methodology

To explore the culturally specific semantic associations of “success” across languages, this study utilizes monolingual word embeddings developed by Joulin et al. [17, 18]. These embeddings, trained on large-scale text corpora from Wikipedia and Common Crawl, represent words as dense vectors in high-dimensional semantic spaces. Crucially, they preserve both semantic similarity and subword information, making them especially effective for languages with rich morphology or less-resourced linguistic contexts.

Each embedding space is language-specific, thereby capturing the nuanced ways in which semantic associations are shaped by the cultural and discursive patterns within that language community [19, 20]. In this study, we select "success" as the anchor concept in each language and compute its semantic proximity to a curated list of culturally significant values and traits (e.g., ambition, effort, wealth, freedom, status, etc.). These concepts were uniformly translated across all target languages to ensure cross-linguistic comparability.

The primary metric for quantifying these associations is the cosine similarity between word vectors. Cosine similarity is a normalized measure that captures the angle between two vectors, reflecting the direction of their semantic orientation irrespective of their magnitude [21]. This property makes cosine similarity particularly suitable for analyzing conceptual relatedness within high-dimensional vector spaces and has been widely applied in sociolinguistic and cultural NLP research [12, 22].

By using these vector-based measurements, closeness in semantic space is treated as a proxy for cultural salience or relevance. For example, if the word for "success" in a given language is closely aligned with "effort" or "wealth," this is interpreted as indicative of a cultural valuation of those traits in relation to the concept of success. This method allows for a scalable, data-driven approach to examining cultural variation without relying on survey instruments or interpretive interviews.

1.3 Novelty and Contribution

This study offers a novel interdisciplinary contribution at the intersection of computational linguistics, cultural sociology, and cross-cultural psychology. While prior research has employed word embeddings to trace historical language change [12], ideological shifts [22] or gender bias [23], few studies have leveraged monolingual embedding spaces as direct proxies for cross-cultural value systems.

Core innovation lies in using static, language-specific embeddings—rather than multilingual or translation-aligned embeddings—as mirrors of cultural meaning-making. Unlike multilingual embeddings that are optimized for cross-language equivalence, monolingual embeddings preserve culture-specific usage patterns [19]. As such, they allow us to infer intra-cultural conceptual hierarchies with greater fidelity.

Furthermore, while previous sociological research on values has relied heavily on surveys and questionnaires—methods which are resource-intensive and vulnerable to self-report bias—used approach offers a scalable, language-based proxy for cultural cognition [24, 25]. By mining large-scale linguistic data, we tap into naturally occurring discourse, capturing the implicit, everyday salience of concepts without needing direct elicitation.

In addition, this study introduces a comparative framework for analyzing the semantic centrality of key success-related values across ten distinct linguistic-cultural contexts: Arabic, Chinese, English, Finnish, German, Hindi, Russian, Spanish, Turkish, and Japanese. To author knowledge, this is one of the first efforts to systematically map semantic models of success across such a wide and diverse range of languages using a unified, quantitative methodology.

2. Literature Review

Recent advancements in computational linguistic methods, particularly through the use of word embeddings, have provided significant insights into the study of cultural narratives. Scholars have utilized these techniques to analyze semantic shifts and word association networks, revealing how language reflects and constructs cultural meanings over time. For instance, Deyne and Storms examined word associations through network and semantic properties, highlighting the small-world structure of word associations and its implications for understanding semantic relationships across cultures [26]. Their findings align with the general consensus that language is inherently social and shaped by the cultural contexts in which it operates.

Similarly, Utsumi's work on distributional semantic models emphasizes how word co-occurrence statistics can effectively model human semantic knowledge, thereby reflecting the influence of cultural frameworks on language use [27]. This aligns with the perspective presented by Steyvers and Tenenbaum, who found that large-scale semantic networks exhibit characteristics of complex, interconnected structures, effectively modeling language acquisition and semantic understanding [28]. Their exploration into the statistical properties of these networks reinforces the notion that cultural dynamics are deeply embedded in linguistic structures.

The notion of semantic facilitation is also underscored by Griffiths et al., who employed network analysis to elucidate the connections within semantic networks, thereby illustrating the ways in which cultural narratives can be quantitatively assessed via computational methods [29]. The evolution of language and its ties to social constructs are evident in studies like those of Hills et al., who tracked the longitudinal analysis of semantic networks, providing insights into how cultural shifts influence word learning and usage over time [30].

Moreover, as demonstrated by Kovács et al. [31], the community structures within word association networks can reveal not only linguistic but also social and cultural categories, illustrating how networks of words reflect broader societal values and norms. Such analyses point to the vital role of network approaches in cultural studies, as they uncover hidden patterns and structures underlying language use, supporting the argument that language serves as a cultural artifact.

The integration of computational linguistic methods with traditional cultural theories represents a fertile ground for generating new insights into language and culture. By combining large-scale text analysis with established theoretical frameworks, researchers have the opportunity to deepen their understanding of cultural phenomena. This interdisciplinary approach leverages the strengths of computational methodologies to address longstanding sociological inquiries regarding cultural narratives, social identities, and collective memory.

For example, Hunter and Smith's application of network text analysis to film narratives demonstrates how linguistic techniques can reveal cultural codes embedded in storytelling, thereby linking computational analysis with sociological insights into genre conventions [32]. The findings from such studies suggest that language not only conveys meaning but also acts as a medium through which cultural values are negotiated and expressed.

Emphasizing collaboration between computational methods and sociological inquiry, researchers like Guo et al. [33], have employed activation force-based measures to analyze complex networks, reflecting the interrelations between language and cognition as influenced by sociocultural factors. By recognizing that word associations and linguistic constructions are contingent upon cultural contexts, scholars can utilize computational techniques to unveil the subtleties of cultural discourse in ways that traditional qualitative methods may not fully capture.

Ultimately, this integration paves the way for methodological innovations that can elucidate the dynamic interplay between language and culture. The combination of quantitative data analysis with theoretical substantiation helps illuminate how cultural constructs evolve and manifest linguistically. Studies utilizing approaches from both realms endorse the belief that computational methods can augment sociological interpretations, yielding richer, multidimensional understandings of how we communicate and construct meaning within varied cultural landscapes.

3. Methodology

This study employs a computational linguistic approach to investigate how different cultures semantically associate the concept of "success" with various related terms. Using monolingual word embeddings for nine languages—Arabic, Chinese, English, Finnish, German, Hindi, Russian, Spanish, and Turkish—we analyze the relative semantic proximity of the word "success" to a set of culturally and conceptually relevant terms.

3.1 Word Selection and Translation

To ensure conceptual consistency across languages, we began with a canonical list of English words representing personal traits, social conditions, and life outcomes frequently associated with success. These included terms such as wealth, effort, education, discipline, luck, and others. For each target language, native-speaking collaborators or validated translation tools were used to generate semantically equivalent terms that reflect typical usage in that cultural and linguistic context. The complete set of words used in this study—including the main word for "success" and the corresponding related terms in each language—is presented in Appendix as Table A1.

3.2 Embedding Models

The analysis utilizes pre-trained monolingual word embeddings released by Joulin et al. [17], which are publicly available. Each language's embeddings are trained independently on respective Common Crawl or Wikipedia corpora using subword information, allowing robust representation even for infrequent words. These embeddings reflect semantic similarity as encoded in the co-occurrence patterns within large-scale text data, which are taken here as proxies for culturally dominant narratives.

To manage computational efficiency, only the vectors for the words of interest were loaded from each language's embedding file. This selective loading strategy significantly reduced memory requirements while maintaining full fidelity for the analysis.

3.3 Semantic Similarity Measurement

The core of the analysis involves computing the cosine distance between the vector representing the main word for "success" and the vectors of the associated terms within each language's embedding space. Cosine distance, defined as one minus the cosine similarity, captures the angular difference between two word vectors and serves as a widely accepted metric for semantic dissimilarity. A smaller cosine distance indicates a closer semantic relationship between the two words in the given language corpus.

By calculating the cosine distances between "success" and each associated term, we are able to construct a profile of conceptual proximity unique to each cultural-linguistic context. These profiles form the basis for both within-culture evaluations and cross-cultural comparisons that follow in the subsequent sections.

4. Results

This study analyzed the semantic associations between the concept of success and fifteen related terms across eight different languages using cosine distance values derived from vector representations. In the context of cosine distances, lower values indicate higher semantic similarity—i.e., a stronger association with the word “success”—while higher values denote weaker associations. The distances obtained as the result of the analyses are given in Table A2 and related heatmap for visualization of data is given in Fig. 1. Darker values indicate closer relation for the heatmap.

Across the eight languages analyzed—Arabic, Chinese, English, Finnish, German, Hindi, Russian, Spanish, and Turkish—some terms such as effort, ambition, and talent consistently showed lower distance values, suggesting a close semantic connection with success. In contrast, words like status, wealth, and respect tended to exhibit higher distances, indicating a weaker association with the concept.

Table 1. Some common patterns for considered words

Term	Common Pattern
Effort	One of the most strongly associated terms across nearly all languages
Ambition	Consistently shows strong semantic connection, especially in Arabic, English
Talent	Often appears among top associated terms in Finnish, Russian, Turkish
Luck	Highly associated in German, Russian, and English, less so in Asian languages
Respect, Status, Wealth	Frequently show weaker associations across most languages

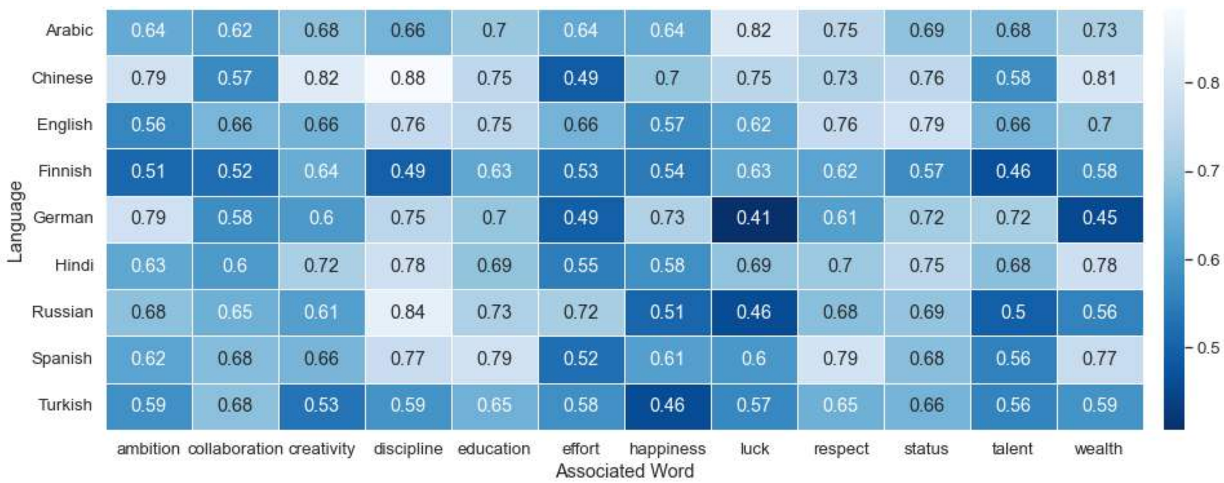


Fig. 1. Heatmap of distance to success of associated words for difference languages.

4.1 Language-Specific Observations

Arabic: Effort (0.6355), ambition (0.6364), and happiness (0.6384) were the most strongly linked with success. In contrast, respect (0.7496), wealth (0.7253), and status (0.6878) were among the least associated.

Chinese: The closest associations appeared with effort (0.4872), collaboration (0.5739), and talent (0.5789). The highest distances—and thus weakest connections—were observed with creativity (0.8249), discipline (0.8845), and power (0.8123).

English: Strongest semantic ties to success were found with ambition (0.5648), happiness (0.5727), and luck (0.6222). Less associated were status (0.7860), power (0.8180), and respect (0.7633).

Finnish: This language showed a close clustering of strong associations, particularly with talent (0.4641), discipline (0.4947), and ambition (0.5104). Weakest associations included education (0.6279), luck (0.6270), and respect (0.6212).

German: Words like luck (0.4060), wealth (0.4541), and effort (0.4942) were most strongly connected to success. The weakest associations were with respect (0.6148), ambition (0.7866), and discipline (0.7487).

Hindi: Closest associations occurred with effort (0.5482), happiness (0.5835), and ambition (0.6285). In contrast, discipline (0.7839), wealth (0.7812), and status (0.7498) showed the least similarity.

Russian: Strongest links were seen with luck (0.4581), talent (0.4955), and happiness (0.5064). Higher distances—and weaker links—were recorded for discipline (0.8427), respect (0.6811), and status (0.6924).

Spanish: Effort (0.5235), talent (0.5574), and ambition (0.6214) had the strongest semantic relations with success, while respect (0.7941), wealth (0.7703), and discipline (0.7741) were less connected.

Turkish: Happiness (0.4565), talent (0.5597), and effort (0.5822) were most closely associated with success. Words like status (0.6570), respect (0.6499), and collaboration (0.6843) showed weaker ties.

5. Discussion

The semantic associations between success and related terms across eight languages reveal both universal and culturally specific patterns. Notably, concepts such as effort, ambition, and talent consistently exhibit strong associations with success, suggesting a shared emphasis on individual merit and personal achievement. Conversely, terms like status, wealth, and respect often show weaker associations, indicating that external markers of success may be less central in certain cultural contexts.

Universal Emphasis on Effort and Talent

Across the languages analyzed, effort emerges as one of the most strongly associated terms with success. This finding reflects the nearly universal endorsement of perseverance and hard work as key contributors to achievement [34]. Similarly, talent shows strong associations, particularly in Finnish, Russian, and Turkish contexts, suggesting that natural aptitude is also widely considered an important driver of success.

These results align with Hofstede's [35] framework, in which cultures with low power distance and high individualism often emphasize personal achievement and meritocracy. However, the consistent emphasis on effort and talent even in more collectivist cultures points to their global importance in perceptions of success.

Cultural Variations in Perceptions of Success

While some patterns are universal, notable cultural differences are evident. For example, luck shows a relatively strong association with success in the German and Russian contexts, indicating a belief in external and uncontrollable factors. This is consistent with prior findings suggesting that people in uncertainty-tolerant cultures may be more willing to attribute outcomes to chance or fate [36].

On the other hand, ambition is strongly associated with success in Arabic and English-speaking cultures. This reflects the cultural emphasis on individual initiative and goal-setting often observed in these societies [37]. In collectivist cultures, like those in parts of Asia, success is often viewed through the lens of communal harmony and collaboration rather than pure individual ambition [38].

Alignment with Cultural Models of Success

The observed semantic associations align with existing cultural models of success. In Latin American cultures, for example, ambition and striving for success are often valued within the bounds of family and community loyalty [39]. Meanwhile, in Chinese culture, the importance of guanxi—social networks and interpersonal connections—helps explain the relatively strong association between success and collaboration in the Chinese dataset [40]. These associations suggest that while effort and talent are broadly appreciated, cultural context significantly shapes how success is perceived, pursued, and achieved.

Implications for Cross-Cultural Understanding

Understanding the cultural semantics of success is critical for global collaboration, international education, and multicultural organizational management. The universal appeal of effort and talent supports merit-based strategies across cultural boundaries. At the same time, recognizing culturally specific success indicators—such as collaboration in Chinese contexts or ambition in Arabic societies—can enhance intercultural empathy and effectiveness [41].

6. Summary & Conclusion

This study employed monolingual Large Language Model embeddings to quantify how the concept of success is semantically related to twelve culturally salient attributes—effort, ambition, talent, collaboration, happiness, luck, creativity, discipline, education, status, wealth, and respect—across nine languages. Lower cosine distances indicate stronger associations with "success," while higher distances denote weaker links. Summary of key findings may be listed as:

- Universal Centrality of Effort: In every language, effort consistently appeared among the closest terms to success (e.g. Chinese 0.487; Spanish 0.524; German 0.494; Finnish 0.529). This underscores a near-universal cultural belief that perseverance and hard work are fundamental to achievement.
- Cultural Variations in Collaboration and Ambition: Collaboration is most tightly bound to success in Finnish (0.522) and Chinese (0.574); and also ranks highly in Arabic (0.616), and English (0.663). This suggests that some cultures, especially those with more collectivist orientations, view cooperative effort as integral to success.
- Ambition shows its strongest ties in Finnish (0.510), English (0.565), and Turkish (0.585), reflecting societies that prize individual drive and goal-setting. In contrast, Chinese (0.788) and Russian (0.685) place less semantic weight on ambition.
- Talent and Happiness as Differentiators: Talent is exceptionally closely linked with success in Finnish (0.464), Russian (0.495), and Spanish (0.557) corpora, indicating that natural aptitude is viewed—sometimes even more than ambition—as a key success ingredient.
- Happiness is most strongly associated in Turkish (0.456), English (0.573), Russian (0.506), and Finnish (0.541), pointing to cultures that incorporate well-being and personal fulfillment into their conceptualization of success.
- Role of Luck in European Contexts: Luck exhibits its lowest cosine distances—and thus strongest semantic ties—in German (0.406) and Russian (0.458) texts, suggesting an attribution of success to chance or external circumstances more than in many Asian or Middle Eastern contexts.

- Weaker Links with External Markers: Across most languages, status and respect tend to have higher distances, indicating that material or socially granted markers are less semantically central to the notion of success than personal qualities and actions.
- The universal prominence of effort aligns with cross-cultural theories that emphasize hard work as a near-universal virtue in goal attainment [34].
- Strong association in Chinese and Arabic contexts accords with these cultures' collectivist orientations [38].
- The high semantic weight of ambition in English and Finnish reflects individualistic value systems, whereas its weaker association in Russian and Chinese embeddings mirrors their more collective or fate-oriented outlooks [35].
- The notable role of luck in German and Russian corpora supports findings that high-uncertainty-avoidance cultures sometimes acknowledge external influences on outcomes [36].

Overall, this computational analysis of word embeddings confirms that while certain values—particularly effort—are universally linked to success, other associations vary in ways that map onto established cultural dimensions. By quantifying these semantic patterns, we gain a nuanced, data-driven perspective on how different societies conceptualize success, complementing traditional sociological and psychological insights. However, this study has certain limitations that should be noted. Specifically, the semantic field analysis is restricted to individual lexical units rather than a broader network incorporating behavioral or experiential data. While this approach allows for precise lexical comparisons, it may not capture dynamic, context-dependent semantic relationships that might emerge in real-world usage or embodied cognition frameworks. Future research could extend this work by integrating behavioral datasets or experimental methods to explore how semantic fields operate within situated contexts.

References

[1] Hofstede G, Hofstede GJ. Dimensions of national cultures. Divers unity cross-cultural Psychol. 1982;173–87.

[2] Inglehart R. *Christian welzel modernization, cultural change, and democracy The human development sequence*. Cambridge: Cambridge university press; 2005.

[3] Göttlich U. Cultural Studies and Sociology of Culture in Germany: Relations and Interrelations. Cult Unbound. 2013; 5(1):33–41.

[4] McGuigan J, Moran M. Raymond Williams and Sociology. Sociol Rev. 2014; 62(1):167–88.

[5] Watts I. Book Review: Lyn Spillman, What Is Cultural Sociology? Sociol Res Online. 2023; 28(4):1160–2.

[6] Su C. Geert Hofstede's Cultural Dimensions Theory and Its Implications in SLA. Acad J Humanit Soc Sci. 2022; 5(14).

[7] Li Y, Kobsa A, Knijnenburg BP, Carolyn Nguyen M-H. Cross-Cultural Privacy Prediction. Proc Priv Enhancing Technol. 2017; 2017(2):113–32.

[8] Kaufmann W, Hooghiemstra R, Feeney MK. Formal institutions, informal institutions, and red tape: A comparative study. Public Adm. 2018; 96(2):386–403.

[9] Taras V, Kirkman BL, Steel P. Examining the impact of Culture's consequences: A three-decade, multilevel, meta-analytic review of Hofstede's cultural value dimensions. J Appl Psychol. 2010; 95(3):405–39.

[10] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013;

[11] Goldberg Y. Neural Network Methods for Natural Language Processing. Synth Lect Hum Lang Technol. 2017; 10(1):1–309.

[12] Hamilton WL, Leskovec J, Jurafsky D. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.

[13] Rutherford G, Tuntivivat S. Social-ecological semiotics and the complex organization of psyche, language, and culture. Cult Psychol. 2024; 30(4):871–86.

[14] Xu J. Brief Analysis on Cross-cultural Communication. Theory Pract Lang Stud. 2011; 1(7).

[15] Ng V, Woo SE, Tay L, Foster J. Examining Variability in Values Attributed to Culture. J Cross Cult Psychol. 2016; 47(7):981–96.

[16] Parajuli B. Role of Language in Shaping Cultural Identity. Marsyangdi J. 2021;112–8.

[17] Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText.zip: Compressing text classification models. arXiv Prepr arXiv161203651. 2016;

[18] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. arXiv Prepr arXiv160701759. 2016;

[19] Hellrich J, Hahn U. Bad company—neighborhoods in neural embedding spaces considered harmful. In: Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers. 2016. p. 2785–96.

[20] Bender EM. On achieving and evaluating language-independence in NLP. Linguist Issues Lang Technol. 2011; 6(3).

[21] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111–9.

[22] Kozlowski AC, Taddy M, Evans JA. The geometry of culture: Analyzing the meanings of class through word embeddings. Am Sociol Rev. 2019; 84(5):905–49.

[23] Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci. 2018; 115(16).

[24] Hofstede G. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications; 1980.

[25] Schwartz SH. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Adv Exp Soc Psychol. 1992; 25:1–65.

[26] De Deyne S, Storms G. Word associations: Network and semantic properties. Behav Res Methods. 2008; 40(1):213–31.

[27] Utsumi A. A Complex Network Approach to Distributional Semantic Models. Zhang Z-K, editor. PLoS One. 2015; 10(8):e0136277.

[28] Steyvers M, Tenenbaum JB. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cogn Sci. 2005; 29(1):41–78.

[29] Griffiths TL, Steyvers M, Firl A. Google and the Mind. Psychol Sci. 2007; 18(12):1069–76.

[30] Hills TT, Maouene M, Maouene J, Sheya A, Smith L. Longitudinal Analysis of Early Semantic Networks. Psychol Sci. 2009; 20(6):729–39.

[31] Kovács L, Bóta A, Hajdu L, Krész M. Networks in the mind – what communities reveal about the structure of the lexicon. Open Linguist. 2021; 7(1):181–99.

[32] Hunter SD, Smith S. A Network Text Analysis of David Ayer's Fury. Adv Lang Lit Stud. 2015; 6(6).

[33] Guo J, Guo H, Wang Z. An Activation Force-based Affinity Measure for Analyzing Complex Networks. Sci Rep. 2011; 1(1):113.

[34] Duckworth AL, Peterson C, Matthews MD, Kelly DR. Grit: Perseverance and passion for long-term goals. J Pers Soc Psychol. 2007; 92(6):1087–101.

[35] Hofstede G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage Publications; 2001.

[36] Spector PE, Cooper CL, Sparks K. An International Study of the Psychometric Properties of the Hofstede Values Survey Module 1994: A Comparison of Individual and Country/Province Level Results. Appl Psychol. 2001; 50(2):269–81.

[37] Furnham A. Personality and learning style: A study of three instruments. Pers Individ Dif. 1992; 13(4):429–38.

[38] Triandis HC. *Individualism & collectivism*. Westview Press; 1995.

[39] Torres L, Driscoll MW, Voell M. Discrimination, acculturation, acculturative stress, and Latino psychological distress: a moderated mediational model. Cult Divers Ethn Minor Psychol. 2012; 18(1):17.

[40] Chen X-P, Chen CC. On the Intricacies of the Chinese Guanxi: A Process Model of Guanxi Development. Asia Pacific J Manag. 2004; 21(3):305–24.

[41] Thomas DC, Inkson KC. *Cultural intelligence: Surviving and thriving in the global village*. Berrett-Koehler Publishers; 2017.

Table A1. The list of used words and their assumed equivalents in considered languages

English	Spanish	German	Japanese	Turkish	Russian	Arabic	Chinese	Finnish	Hindi
success	éxito	erfolg	成功	başarı	успех	نجاح	成功	menestys	सफलता
money	dinero	geld	お金	para	деньги	مال	钱	raha	धन
status	estatus	status	地位	statü	статус	مكانة	地位	asema	दर्जा
effort	esfuerzo	anstrengung	努力	çaba	усилие	جهد	努力	ponnistus	प्रयास
ambition	ambición	ambition	志	hırs	амбиция	طموح	抱负	kunnianhimo	महत्वाकांक्षा
collaboration	colaboración	zusammenarbeit	協力	işbirliği	сотрудничество	تعاون	合作	yhteistyö	सहयोग
education	educación	bildung	教育	eğitim	образование	تعليم	教育	opetus	शिक्षा
talent	talento	talent	才能	yetenek	талант	موهبة	才能	lahjakkuus	प्रतिभा
creativity	creatividad	kreativität	創造力	yaratıcılık	креативность	إبداع	创造力	luovuus	रचनात्मकता
luck	suerte	glück	幸運	şans	удача	حظ	运气	tuuri	भाग्य
discipline	disciplina	disziplin	規律	disiplin	дисциплина	انضباط	纪律	kurinalaisuus	अनुशासन
happiness	felicidad	Glück	幸福	mutluluk	счастье	سعادة	幸福	onnellisuus	खुशी
respect	respeto	respekt	尊敬	saygı	уважение	احترام	尊重	kunnioitus	सम्मान

Table A2. The cosine distance results for the considered words and languages

Language	Considered Word	Cosine Distance	Language	Considered Word	Cosine Distance
Arabic	collaboration	0.616	Hindi	effort	0.548
Arabic	effort	0.636	Hindi	happiness	0.584
Arabic	ambition	0.636	Hindi	collaboration	0.604
Arabic	happiness	0.638	Hindi	ambition	0.628
Arabic	discipline	0.658	Hindi	talent	0.682
Arabic	talent	0.678	Hindi	luck	0.689
Arabic	creativity	0.684	Hindi	education	0.691
Arabic	status	0.688	Hindi	respect	0.703
Arabic	education	0.698	Hindi	creativity	0.724
Arabic	wealth	0.725	Hindi	status	0.750
Arabic	respect	0.750	Hindi	wealth	0.781
Arabic	luck	0.815	Hindi	discipline	0.784
Chinese	effort	0.487	Russian	luck	0.458
Chinese	collaboration	0.574	Russian	talent	0.495
Chinese	talent	0.579	Russian	happiness	0.506
Chinese	happiness	0.696	Russian	wealth	0.559
Chinese	respect	0.734	Russian	creativity	0.613
Chinese	luck	0.746	Russian	collaboration	0.650
Chinese	education	0.754	Russian	respect	0.681
Chinese	status	0.755	Russian	ambition	0.685
Chinese	ambition	0.788	Russian	status	0.692
Chinese	wealth	0.808	Russian	effort	0.719
Chinese	creativity	0.825	Russian	education	0.731
Chinese	discipline	0.885	Russian	discipline	0.843
English	ambition	0.565	Spanish	effort	0.524
English	happiness	0.573	Spanish	talent	0.557
English	luck	0.622	Spanish	luck	0.598
English	creativity	0.655	Spanish	happiness	0.612
English	effort	0.658	Spanish	ambition	0.621
English	talent	0.658	Spanish	creativity	0.661
English	collaboration	0.663	Spanish	collaboration	0.678
English	wealth	0.695	Spanish	status	0.680
English	education	0.750	Spanish	wealth	0.770
English	respect	0.763	Spanish	discipline	0.774
English	discipline	0.765	Spanish	education	0.789
English	status	0.786	Spanish	respect	0.794
Finnish	talent	0.464	Turkish	happiness	0.456
Finnish	discipline	0.495	Turkish	creativity	0.531

Language	Considered Word	Cosine Distance	Language	Considered Word	Cosine Distance
Finnish	ambition	0.510	Turkish	talent	0.560
Finnish	collaboration	0.522	Turkish	luck	0.567
Finnish	effort	0.529	Turkish	effort	0.582
Finnish	happiness	0.541	Turkish	ambition	0.585
Finnish	status	0.575	Turkish	wealth	0.590
Finnish	wealth	0.584	Turkish	discipline	0.591
Finnish	respect	0.621	Turkish	respect	0.650
Finnish	luck	0.627	Turkish	education	0.651
Finnish	education	0.628	Turkish	status	0.657
Finnish	creativity	0.642	Turkish	collaboration	0.684
German	luck	0.406	German	education	0.696
German	wealth	0.454	German	status	0.715
German	effort	0.494	German	talent	0.720
German	collaboration	0.579	German	happiness	0.734
German	creativity	0.601	German	discipline	0.749
German	respect	0.615	German	ambition	0.787

Research & Design
www.jrede.org

Research Article

The integration of the Hook Model into E-sports marketing: A promotional perspective

Sunay Güngör Gülsoy*[a]

Department of Business, Süleyman Demirel University, Isparta, Türkiye.

Article Info

Article History:

Received:
23 May 2025

Accepted:
22 June 2025

Keywords:

Behavioral design;
Digital marketing;
E-Sports;
Hook Model;
Promotion strategy;
User engagement

Abstract

This paper explores the integration of the Hook Model—comprised of the stages Trigger, Action, Variable Reward, and Investment—into e-sports marketing, with a particular focus on the promotional component of the marketing mix. As e-sports emerges as a dominant force within the digital entertainment industry, traditional marketing approaches are being replaced by behaviorally driven frameworks designed to cultivate user habits and enhance engagement. Through a detailed examination of the e-sports ecosystem—including developers, players, teams, and spectators—this study analyzes how promotional strategies can be optimized using the Hook Model. Real-world examples, such as time-sensitive rewards and engagement triggers in games like Valorant, demonstrate the practical alignment of this model with audience retention strategies. While the framework proves effective in sustaining user attention and fostering brand loyalty, ethical concerns regarding behavioral manipulation and digital addiction are critically discussed. The study concludes that, when applied responsibly, the Hook Model can serve as a powerful tool for marketers seeking to build enduring connections with e-sports audiences. Future research directions are proposed to further examine the long-term effects and ethical dimensions of habit-forming marketing strategies in digital environments.

1. Introduction

The rapid proliferation of digital technologies has catalyzed the transformation of many traditional sectors, and among the most prominently affected is the sports industry. In recent years, electronic sports (e-sports) have emerged not only as a competitive and recreational activity but also as a significant economic and cultural phenomenon within the digital ecosystem. Defined as organized video game competitions often broadcast to spectators via online platforms, e-sports encapsulate a hybrid identity combining elements of gaming, media, entertainment, and professional sport [1]. With millions of participants and an ever-expanding audience base, e-sports have evolved into a global industry supported by an elaborate infrastructure of game developers, players, teams, fans, and sponsors [2].

This growth has been paralleled by a strategic shift in marketing practices, particularly in how brands interact with and retain digital audiences. The promotion component of the marketing mix (4Ps); encompassing advertising, public relations, direct marketing, sponsorships, and sales promotions has become increasingly critical in the e-sports

*Corresponding author: sunaygulsoy@sdu.edu.tr
[a]orcid.org/0000-0001-8390-9030
DOI: http://dx.doi.org/10.17515/rede2025-005ec0523rs

landscape. Unlike traditional advertising models, e-sports marketing requires an adaptive, immersive, and behaviorally-informed approach to capture the attention and loyalty of digital-native users. It is within this context that the Hook Model, developed by Nir Eyal and Ryan Hoover (2014), offers an intriguing theoretical and practical framework [3].

The Hook Model presents a four-phase loop—Trigger, Action, Variable Reward, and Investment—designed to cultivate user habits and foster long-term engagement with digital products. While originally conceived for app development and behavioral product design, the model holds promising implications for digital marketing strategies, especially within environments such as e-sports where user engagement and behavioral reinforcement are central to both gameplay and brand interaction.

This paper aims to explore the integration of the Hook Model into e-sports marketing, particularly through the lens of the promotion element of the marketing mix. By mapping the model's components onto real-world promotional tactics used in the e-sports industry, the study endeavors to offer a novel interpretive framework for understanding and enhancing digital audience engagement. A key premise of this inquiry is that the Hook Model, when effectively embedded into marketing strategies, can provide a systematic method for increasing user retention, fostering community participation, and ultimately amplifying brand loyalty within the e-sports domain.

To this end, the subsequent sections of the article will provide a conceptual overview of the e-sports ecosystem and its stakeholders, outline the theoretical underpinnings of the Hook Model, demonstrate the model's practical applicability within e-sports promotional strategies—drawing on examples such as Valorant—and engage in a critical discussion of the model's effectiveness, limitations, and implications for the future of digital sports marketing.

2. Literature Review

The The intersection of e-sports, digital marketing, and behavioral design has attracted growing scholarly attention, driven by the convergence of entertainment technologies and participatory media cultures. To establish a comprehensive understanding of the topic, this section examines core areas of literature: the e-sports ecosystem, promotional strategies in digital marketing, and the theoretical foundations of the Hook Model.

E-sports, once considered a niche subculture, has matured into a structured industry encompassing multiple stakeholders including game developers, players, professional teams, audiences, and sponsors. The ecosystem is anchored by competitive video games such as League of Legends, Valorant, and Counter-Strike, which provide the stage for organized tournaments and broadcasting events [1]. Game developers, in particular, hold significant control over content creation and monetization pathways. As Yükçü and Kaplanoğlu (2018) argue, these companies have evolved from mere content providers into multi-faceted digital entertainment corporations that integrate social media, influencer marketing, and community engagement into their core strategies [4].

The rise of mobile gaming, broadband access, and video streaming platforms like Twitch and YouTube Gaming has further intensified the reach and cultural relevance of e-sports. Akgöl (2019) notes that the industry's financial model is largely sustained by audience-driven revenue streams, including advertising, subscriptions, and sponsorships. These dynamics underscore the importance of understanding audience behavior and the mechanisms by which engagement can be both initiated and sustained [5].

In the context of e-sports, promotion—one of the 4Ps in the traditional marketing mix—is not just about advertising, but rather about creating immersive brand experiences.

Promotional strategies typically include live-streamed sponsorships, influencer partnerships, social media campaigns, virtual product placements, and gamified promotions [6]. The boundary between player, viewer, and consumer is often blurred, as audiences participate in real-time chats, fan economies, and user-generated content that feed back into the promotional ecosystem.

Hutchins (2006) emphasizes that promotional content in e-sports is designed not only to inform but also to cultivate ritualistic user behavior. Unlike static media advertisements, e-sports promotions are dynamic and often embedded directly into the content of gameplay or stream narratives, making them more effective in shaping viewer perception and driving consumer action. This environment provides fertile ground for behavioral design frameworks such as the Hook Model to be operationalized [7].

Originally introduced in the field of product design, the Hook Model proposed by Eyal (2014) outlines a cyclical process that facilitates habit formation through four phases: Trigger, Action, Variable Reward, and Investment. Triggers—either external (e.g., notifications) or internal (e.g., emotional needs)—initiate user behavior. This is followed by a simple action, a variable reward that satisfies the user's need, and a subsequent investment that increases the likelihood of repeating the cycle. The model has been widely adopted in the design of mobile apps, social media platforms, and digital services [3]. While empirical applications of the Hook Model have predominantly focused on user interface design, its relevance to digital marketing strategies has begun to emerge. Fogg (2009), whose Behavioral Model for Persuasive Design inspired the Hook Model, underscores the critical interplay between motivation, ability, and triggers in shaping user behavior. When viewed through a marketing lens, the Hook Model provides a framework for understanding how promotional strategies can foster habitual brand interactions, particularly in high-engagement environments like e-sports [8].

However, critiques of the model point to ethical concerns surrounding user manipulation and digital addiction, particularly in industries designed to maximize screen time and consumption [9]. This tension is especially salient in the e-sports context, where the user experience is gamified by design, and where promotional content is deeply entwined with entertainment.

3. The E-Sports Ecosystem and Its Stakeholders

The e-sports ecosystem is a complex, multi-layered network composed of various stakeholders who interact dynamically within a digital and often gamified environment. Understanding the functions and motivations of these stakeholders is essential for contextualizing how promotional strategies and behavioral engagement models, such as the Hook Model, can be effectively implemented. This section delineates the principal actors in the e-sports landscape: game developers, video games as platforms, individual players and teams, and spectators. Each plays a vital role in shaping the economic and cultural logic of the industry.

3.1 Game Developers

Game developers serve as the foundational pillars of the e-sports ecosystem. They are responsible not only for the creation and technical maintenance of e-sports titles but also for curating the broader user experience, including in-game economies, competitive structures, and digital community spaces. As the industry has matured, developers have had to transition from a product-centric to an experience-centric model of operation [4]. No longer limited to designing software, developers now strategically leverage digital marketing channels, sponsorships, and influencer partnerships to create emotionally resonant and habit-forming experiences. For instance, Riot Games, the developer of League of Legends and Valorant, invests heavily in community engagement through seasonal events, limited-time content, and integrated promotional triggers—mechanisms that align with the Hook Model's emphasis on frequent and behaviorally potent interaction cycles.

3.2 Video Games as Interactive Platforms

Video games within the e-sports domain function not just as leisure activities but as immersive media platforms. Their design incorporates interactive features such as real-time feedback, customizable avatars, progression systems, and social features—all of which serve as mechanisms of both user engagement and monetization [10]. These platforms act as living ecosystems where promotional activities are seamlessly embedded into gameplay experiences. In doing so, video games themselves become fertile grounds for behavior-shaping models such as the Hook Model to take root. For example, time-limited in-game promotions often trigger user engagement (Trigger), elicit a user action (Action), offer a randomized reward (Variable Reward), and encourage future investment in the form of in-game currency or time (Investment).

3.3 Players and Teams

The role of players—both individual and team-based—is central to the e-sports industry. Players are not merely consumers of digital content but active participants who co-create the e-sports experience through their performances, interactions, and social media presence. Competitive players train rigorously, often in structured environments that mirror those of traditional sports teams, complete with coaches, analysts, and sponsorship contracts [1]. These players also serve as brand ambassadors and content creators, playing a dual role that merges competition with marketing. Their participation in events, product endorsements, and social streams often initiates Triggers in the Hook Model—encouraging fans and casual players alike to engage with promoted content or return to gameplay environments.

3.4 Spectators and the Audience Economy

Spectators constitute the largest and arguably most economically vital component of the e-sports ecosystem. Their participation, while often passive in terms of gameplay, is highly active in terms of content consumption, community engagement, and monetization. As Akgöl (2019) emphasizes, e-sports is a "bottom-up" movement, heavily reliant on the interactive nature of platforms such as Twitch, YouTube Gaming, and Discord. These environments facilitate parasocial relationships, fan economies, and user-generated content, all of which amplify the reach and effectiveness of promotional efforts. Spectators are frequently subjected to external triggers—such as live notifications, influencer cues, or limited-time viewing rewards—that entice them into cycles of habitual engagement [5].

Moreover, spectators play a crucial role in the data economy of e-sports. Their clicks, view times, and spending behaviors are meticulously tracked and analyzed, allowing companies to optimize their promotional strategies using behavioral data. This feedback loop enhances the applicability of the Hook Model by enabling marketers to personalize and iterate on their engagement techniques, thereby reinforcing cycles of attention and loyalty.

4. The Hook Model Explained

The Hook Model, developed by Nir Eyal and Ryan Hoover (2014), is a behavioral design framework intended to explain how digital products can create user habits by repeatedly engaging users through a four-phase cycle: Trigger, Action, Variable Reward, and Investment. Originally applied in the fields of app development and user experience design, this model offers a compelling lens through which to understand the mechanisms of digital marketing—particularly in the context of e-sports, where user engagement is central to the ecosystem's sustainability. This section outlines each stage of the model, with conceptual depth and its relevance to marketing strategies in e-sports environments [3].

4.1 Trigger

Triggers are the initial cues that prompt a user to take action. These can be external, such as notifications, emails, advertisements, or social media alerts, or internal, arising from psychological states like boredom, loneliness, or a need for self-affirmation.

- External Triggers in e-sports include push notifications about upcoming tournaments, limited-time in-game offers, or promotional emails from game developers. For example, a message like "Valorant Night Market is now live!" functions as an external trigger designed to draw players into the game environment.
- Internal Triggers operate on emotional associations. A user feeling isolated might be internally triggered to open Twitch and watch their favorite streamer, seeking social connection or belonging [11].

These triggers are essential in e-sports marketing, as they initiate the behavioral loop that leads to user re-engagement with the game or content platform.

4.2 Action

The action phase is the simplest behavior performed by the user in anticipation of a reward. For a behavior to occur, the user must have sufficient motivation, ability, and a strong enough trigger—an idea rooted in Fogg's Behavior Model [8].

In the e-sports context, typical user actions include:

- Logging into a game following a trigger.
- Watching a live match on Twitch.
- Commenting on a gameplay video.
- Clicking on a promotional link to redeem an in-game bonus.

The ease and frequency of these actions are designed to require minimal effort, reinforcing user engagement. For instance, developers often simplify login procedures or offer one-click access to promotional events to lower the barrier to entry.

4.3 Variable Reward

The variable reward phase is central to habit formation. Unlike fixed rewards, variable rewards introduce uncertainty and anticipation, making them more psychologically compelling. This concept draws on principles from behavioral psychology, particularly B.F. Skinner's work on operant conditioning.

In e-sports, variable rewards are widely used and take several forms:

- Loot boxes, seasonal chests, or daily login bonuses that offer randomized items.
- Receiving exclusive skins or emotes by watching tournament streams.
- Audience raffles or "drops" during live broadcasts.
- Earning in-game currency after winning competitive matches.

These rewards create dopaminergic anticipation—the uncertainty of what the user might receive drives them to repeat the action. Table 1 illustrates how variable rewards function in both gameplay and marketing domains within e-sports.

Table 1. Examples of Variable Rewards in E-Sports Context

Context	Variable Reward Type	Example
Gameplay	Randomized loot	Weapon skins from cases in Counter-Strike: Global Offensive
Streaming	Viewer drops	Special item drop for watching a League of Legends final
Promotions	Limited-time reward	Discount coupon for purchasing in-game currency
Community	Social feedback	Likes or shoutouts from popular streamers

4.4. Investment

The final stage, investment, involves the user putting something of value—time, effort, data, money—into the product or experience. Investment increases the likelihood of future engagement because the user builds a psychological attachment to their progress or identity within the system.

In e-sports, this is seen when players:

- Level up their in-game avatars or characters.
- Purchase battle passes or exclusive cosmetics.
- Contribute to forums or fan communities.
- Train and strategize with teams to improve competitive rankings.

These investments are not merely sunk costs; they create a sense of ownership and commitment, making users more likely to return to the game or platform and re-enter the Hook cycle. This aligns with IKEA effect principles in behavioral economics, where individuals place higher value on self-crafted experiences [12].

5. Mapping the Hook Model to E-Sports Promotion Strategies

The integration of the Hook Model into e-sports promotion strategies offers a powerful framework for cultivating sustained user engagement through iterative behavioral loops. Within the traditional 4Ps of marketing, promotion encompasses activities such as advertising, public relations, sales promotion, and direct marketing—each designed to inform, persuade, and remind target audiences. In the e-sports industry, where audience attention is fragmented and highly competitive, traditional promotional tactics alone are insufficient. Instead, brands and developers increasingly rely on behaviorally-informed promotion, where psychological engagement models like the Hook Model are embedded into campaign design. This section examines how each stage of the Hook Model can be operationalized through specific promotional mechanisms in the e-sports domain.

5.1. Triggers in Promotional Channels

Promotional triggers in e-sports are frequently designed to capture user attention and initiate a behavioral sequence. These are executed through a combination of external cues (e.g., advertisements, event notifications, emails) and contextually embedded messages within games or streaming platforms.

- Advertising campaigns for in-game events are distributed across platforms like YouTube, Twitch, and Discord.
- Push notifications for promotions (e.g., “Double XP Weekend!”) prompt re-engagement.

- Social media countdowns to new seasons or events activate anticipation and emotional readiness.

These triggers align with the initial phase of the Hook Model, encouraging users to begin a new cycle of interaction with the brand or platform.

5.2. User Action and Ease of Engagement

Once triggered, the next step is to facilitate low-effort user actions. In e-sports promotions, the key is to make the desired behavior—such as watching a stream, entering a code, or joining a limited-time event—frictionless and rewarding.

- Direct engagement includes participating in beta tests, entering sweepstakes, or clicking ads for discounts.
- Micro-interactions such as following a sponsor on social media or registering for an online tournament act as simple but effective behavioral actions.

By reducing the cognitive or procedural cost of action, marketers increase the likelihood that users will complete the behavior, as posited in Fogg's Behavior Model [8].

5.3. Variable Rewards in Promotional Design

To maintain engagement, promotional efforts often include variable reward mechanisms, which introduce uncertainty and create a feedback loop of expectation and reinforcement. These rewards are not only tied to gameplay but also directly to promotional activities.

Table 2. Examples of Variable Rewards in E-Sports Promotion

Promotional Activity	Variable Reward Mechanism	Example
Watching Sponsored Streams	Random viewer rewards (“drops”)	Exclusive in-game items for watching Valorant Champions Tour
Engaging with Social Content	Entry into randomized giveaways	Limited-edition merchandise raffles via retweet campaigns
Purchasing Promotional Bundles	Mystery bonuses or bonus points	Bonus loot boxes after purchase during seasonal events
Participating in Surveys	Entry into lucky draws	Survey completions rewarded with chance to win gaming accessories

This approach leverages the brain’s sensitivity to variable outcomes, keeping users psychologically invested and reinforcing promotional effectiveness.

5.4. Investment through Community and Customization

The final phase, investment, is encouraged when promotional strategies allow users to commit resources—time, money, effort—thereby increasing their emotional and behavioral stake in the ecosystem. In e-sports, investment mechanisms are often designed to build brand attachment and encourage repeated participation.

- Progressive event systems (e.g., battle passes tied to promotional content) reward users who engage consistently over time.
- Exclusive badges, tiers, or avatars associated with campaign participation (e.g., early backer skins or limited-time branding elements) promote identity alignment.

- User-generated content contests, sponsored by brands or game developers, allow fans to contribute and showcase their creativity, increasing perceived ownership.

These investment-driven tactics result in more committed user bases who are more likely to re-enter the Hook loop, ensuring the longevity of the promotional strategy and the user's relationship with the brand or product.

5.5. Integrated Framework

Table 3 below provides a summary of how promotional elements in e-sports can be mapped to the four stages of the Hook Model.

Table 3. Hook Model Aligned with E-Sports Promotion Strategies

Hook Model Phase	E-Sports Promotion Example
Trigger	Notifications about limited-time tournaments
Action	Clicking a link to watch a livestream or redeem a game code
Variable Reward	Randomized rewards for participation or viewing
Investment	Purchasing a bundle, creating content, or progressing a battle pass

By embedding the Hook Model into e-sports promotional strategies, marketers are not only advertising a product or service—they are constructing habitual interaction architectures that blur the boundaries between gameplay, brand engagement, and community belonging.

6. Results and Discussion

The application of the Hook Model in e-sports promotional strategies provides an analytically robust and behaviorally intuitive framework for understanding and enhancing user engagement. However, while the model offers a coherent logic for structuring promotional efforts around habit formation, its real-world implementation within e-sports ecosystems invites both strategic promise and critical reflection. This section analyzes the effectiveness, limitations, and ethical implications of integrating the Hook Model into e-sports marketing, while also considering its broader impact on user behavior, brand loyalty, and the digital consumer experience.

6.1. Effectiveness in Driving Engagement and Retention

One of the key strengths of the Hook Model lies in its capacity to increase user retention by tapping into psychological drivers of behavior. In the competitive attention economy of e-sports, the capacity to keep users "hooked" is essential. Games like Valorant and League of Legends do not merely rely on gameplay quality; they are supported by layered promotional ecosystems that use triggers (e.g., event alerts), encourage simple actions (e.g., log-ins), offer variable rewards (e.g., loot boxes, viewer drops), and demand investment (e.g., time, customization) [11].

This cycle not only strengthens the user's engagement with the game but also deepens their relationship with peripheral entities—streamers, sponsors, and event organizers—thereby increasing brand touchpoints. Empirical evidence from marketing studies suggests that increased exposure and behavioral repetition enhance brand familiarity and positive sentiment [13]. Therefore, the Hook Model's design logic dovetails effectively with the goals of e-sports marketers aiming to transform casual participants into loyal brand advocates.

6.2. Scalability and Flexibility of the Model

Another advantage of the Hook Model is its scalability across diverse promotional formats and user segments. Whether through micro-interactions on social media, gamified loyalty programs, or large-scale tournament campaigns, each stage of the Hook cycle can be tailored to suit the specific needs and behavioral profiles of different e-sports demographics.

For example, casual players may be more responsive to frequent external triggers and small variable rewards (e.g., daily login bonuses), while competitive players might be more engaged through higher-stakes investments such as ranked tournaments or exclusive in-game assets. This adaptability allows developers and marketers to segment audiences and fine-tune promotional content accordingly, maximizing both reach and relevance.

6.3. Ethical Concerns and Behavioral Manipulation

Despite its strategic utility, the Hook Model also raises ethical questions, particularly regarding its potential to exploit cognitive biases and foster compulsive usage. The use of variable rewards—especially when linked to monetized mechanisms like loot boxes—has been criticized for mimicking the psychological triggers of gambling [14]. Critics argue that such systems risk encouraging excessive spending and unhealthy usage patterns, especially among younger or more vulnerable users.Furthermore, the investment phase may lead users to overcommit to a digital ecosystem, making it psychologically difficult to disengage without losing perceived value or status. This phenomenon, related to the sunk cost fallacy, reinforces behaviors that may not always align with users' best interests, particularly when combined with social pressure or fear of missing out (FOMO).These concerns necessitate a responsible marketing approach, wherein the Hook Model is applied with transparency and user well-being in mind. Game developers and marketers must balance commercial objectives with ethical design principles, incorporating opt-out mechanisms, spending limits, and informative disclosures about reward probabilities and promotional content.

6.4. Implications for Future E-Sports Marketing

The successful application of the Hook Model in e-sports marketing highlights a broader shift toward behavior-driven design in digital strategy. As user data becomes more granular and accessible, marketing efforts will increasingly rely on predictive analytics and personalization algorithms to tailor each stage of the Hook cycle to individual users in real-time [15].

This direction opens new possibilities for automated engagement systems that dynamically adapt triggers, rewards, and investment paths based on user behavior. However, it also intensifies the ethical stakes, requiring the development of regulatory frameworks and industry standards to prevent exploitative practices.

Moreover, the integration of the Hook Model invites further academic research. Longitudinal studies could investigate how long users remain in the engagement loop, what triggers lead to disengagement, and how variations in reward structure influence long-term brand perception. Exploring cross-cultural applications of the model could also yield insights into how different communities respond to behavioral marketing strategies within the global e-sports landscape.

5. Conclusions

The application of the Hook Model within the realm of e-sports marketing represents a significant shift in how digital engagement strategies are conceptualized, executed, and evaluated. As demonstrated throughout this study, the four interconnected stages of the Hook Model—Trigger, Action, Variable Reward, and Investment—offer a psychologically grounded and operationally scalable framework for driving sustained user engagement across the e-sports ecosystem.

From the perspective of promotion, one of the core components of the marketing mix, the integration of the Hook Model facilitates more than traditional audience outreach; it supports the creation of habitual user loops that transform spectators into participants, and casual players into brand-loyal community members. By embedding promotional triggers into gameplay mechanics, aligning simple user actions with reward structures, and designing systems that incentivize long-term investment, e-sports organizations and marketers are able to foster deeper and more enduring forms of engagement.

The empirical and conceptual evidence presented suggests that this approach can increase brand affinity, enhance retention, and amplify the emotional resonance of digital campaigns. However, this success comes with clear ethical responsibilities. As the model intentionally leverages cognitive biases and psychological motivators, its misuse can lead to behavioral manipulation, excessive screen time, or compulsive monetization—issues particularly salient in a young and digitally immersed user base. Therefore, responsible implementation—guided by ethical design standards, transparency, and user consent—is paramount.

In addition, the Hook Model's utility extends beyond tactical campaign execution; it encourages a paradigm shift in marketing strategy, wherein behavioral science and digital design coalesce to create ecosystems that are not merely consumed but actively lived in. For the e-sports industry, which thrives on interactivity, community, and user participation, this model provides a lens through which to architect experiences that are both commercially effective and psychologically engaging.

Future research might explore the long-term psychological effects of prolonged engagement through Hook-based systems, cross-cultural variations in behavioral response to such promotional tactics, and comparative effectiveness across different genres or platforms in the e-sports landscape. As e-sports continues to mature and diversify, the demand for data-informed yet ethically conscious promotional strategies will only grow.

Ultimately, the fusion of the Hook Model with e-sports marketing marks an important development in the study of digital consumer behavior one that necessitates continuous reflection, innovation, and accountability.

References

[1] Hamari, J., Sjöblom, M. What is eSports and why do people watch it? Internet Research, 2017; 27-2: 211–232.

[2] Karhulahti, V. M. Reconsidering e-sport: Economics and executive ownership. Physical Culture and Sport. Studies and Research, 2017; 74: 43–53.

[3] Eyal, N. Hooked: How to build habit-forming products. Penguin Uk, 2014.

[4] Yükçü, S., Kaplanoğlu, E. E-spor endüstrisi. Uluslararası İktisadi ve İdari İncelemeler Dergisi, 2018; 21: 533–550.

[5] Akgöl, O. Spor endüstrisi ve dijitalleşme: Türkiye'deki e-spor yapılanması üzerine bir inceleme. TRT Akademi, 2019; 4-2: 206–224.

[6] Seo, Y. Electronic sports: A new marketing landscape of the experience economy. Journal of Marketing Management, 2013; 29: 1542–1560.

[7] B. Hutchins. 'Tales of the Digital Sublime: Tracing the Relationship Between Big Data and Professional Sport', Convergence: The International Journal of Research into New Media Technologies, 2016; 22-5: 494-509.

[8] Fogg, B. J. A behavior model for persuasive design. Proceedings of the 4th International Conference on Persuasive Technology, 40, 2010.

[9] Alter, A. Irresistible: The rise of addictive technology and the business of keeping us hooked. Penguin, 2018.

[10] Jackson, L. A., Games, A. I. Video games and creativity. In A. I. Games & L. A. Jackson (Eds.), Video Games and Creativity (pp. 3–38). Academic Press, 2015.

[11] Eyal, N. . Hooked: How to build habit-forming products. Penguin Uk, 2017.

[12] Norton, M. I., Mochon, D., Ariely, D. The IKEA effect: When labor leads to love. Journal of Consumer Psychology, 2012; 22-3: 453–460.

[13] Reijmersdal, E. A. van, Jansz, J., Peters, O., & Van Noort, G. (2013). Why girls go pink: Game character identification and game-players' motivations. Computers in Human Behavior, 29(6), 2640–2649.

[14] King, D. L., Delfabbro, P. H. (2019). Video game monetization (e.g., 'loot boxes'): A blueprint for practical social responsibility measures. International Journal of Mental Health and Addiction, 17(1), 166–179.

[15] Morgan, C. C. Trauma fractures and the brutality of sport: Exploring evidence of gaming in the Mississippian Southeast (Doctoral dissertation, University of Alabama Libraries), 2016.

Research & Design
www.jrede.org

Research Article

Gamifying learning through escape room design: Educational potential, puzzle typologies, and technology integration

Ceren Demir*[a]

Department of Public Administration, Muğla Sıtkı Koçman University, Muğla, Türkiye.

Article Info

Article History:

Received:
23 May 2025

Accepted:
25 June 2025

Keywords:

Gamification,
Experiential learning,
Educational technology,
Puzzle typology,
Escape room.

Abstract

This conceptual study explores the educational potential of escape room games through a typological analysis of puzzle design. By categorizing puzzles into mental, physical, and meta types, the research outlines how each modality supports distinct cognitive, social, and affective skills-ranging from logical reasoning and spatial manipulation to collaboration and metacognition. Drawing on gamification theory, constructivist learning principles, and cognitive development frameworks, the paper positions escape rooms as transdisciplinary tools with applications in classroom instruction, corporate training, and therapeutic settings. The integration of emerging technologies such as Augmented Reality, Virtual Reality, and IoT-based systems further expands the scope and scalability of escape rooms in hybrid and digital learning environments. Key design considerations-such as puzzle clarity, logical progression, accessibility, and fail-safe mechanisms-are discussed as foundational to both educational effectiveness and user experience. Ultimately, this study argues that escape rooms represent a compelling pedagogical paradigm for 21st-century education, supporting lifelong learning through immersive, game-based experiences.

1. Introduction

Puzzles have occupied a unique place in human history as structured challenges that test reasoning, memory, and logic. Originating in oral traditions and later transitioning into printed forms with the advent of mass media, puzzles have long been recognized not only as recreational tools but also as mediums for intellectual engagement [1]. In Turkey, the first known printed puzzle appeared in 1925 in the Resimli Mecmua, illustrating the early integration of cognitive games into literary and cultural contexts [2]. Over time, puzzles diversified into types such as crosswords, word-image associations, logic grids, and number-based games like Sudoku. Each type, despite differences in form and complexity, is designed to engage the solver in a process of problem recognition, hypothesis formulation, and solution testing-functions that align with core cognitive operations.

Escape rooms have emerged as a dynamic evolution of this puzzle tradition, integrating various types of challenges within immersive and often narrative-driven environments. Originating in Japan and the United States in the early 2000s, escape rooms quickly gained international popularity, becoming both a recreational trend and an innovative pedagogical tool [3]. These experiences typically require participants to solve a series of puzzles within a time limit to "escape" from a themed room. Unlike traditional puzzles, which are often solitary and two-dimensional, escape room puzzles are multi-sensory,

Corresponding author: demircerennn@gmail.com
[a]orcid.org/0000-0001-8721-8261
DOI: http://dx.doi.org/10.17515/rede2025-006ga0523rs

collaborative, and sequentially structured within a game loop that includes clues, solutions, feedback, and rewards. Their application has expanded beyond entertainment into corporate training, therapeutic interventions, and-most notably-formal education [4]. This paper addresses a central research problem: How can different puzzle types in escape rooms enhance cognitive development and educational outcomes? While existing literature has examined escape rooms as engaging tools for learning [4, 5]. less attention has been paid to the typological structure of puzzles and how these contribute to specific learning domains such as critical thinking, spatial reasoning, and teamwork. This is particularly relevant as escape rooms become integrated into curricula and professional development programs, where the educational potential of different puzzle types-mental, physical, and meta-demands critical scrutiny.

The purpose of this conceptual study is to categorize and analyze the various puzzle types used in escape rooms and to explore their contributions to cognitive, affective, and collaborative learning processes. The paper also aims to highlight design considerations and suggest ways in which these puzzles can be adapted or optimized for pedagogical contexts. As such, the significance of this research lies in its potential to inform educators, instructional designers, and game developers about how puzzle mechanics can be intentionally aligned with educational objectives.

This paper is guided by the following research questions:

- What are the defining characteristics of mental, physical, and meta puzzles used in escape rooms?
- How do these puzzle types facilitate cognitive engagement, problem-solving, and knowledge retention?
- In what ways can puzzle design be optimized to align with pedagogical goals and learner variability?
- How can emerging digital technologies extend the applicability and effectiveness of escape room-based puzzles in education?

By addressing these questions, this study contributes to the growing body of literature at the intersection of gamification, cognitive science, and instructional design, while offering a practical framework for educators seeking to implement escape room pedagogy in diverse learning environments.

2. Literature Review

Gamification has emerged as a widely studied framework for enhancing motivation, engagement, and user experience in non-game contexts. Deterding, Dixon, Khaled, and Nacke [6] define gamification as "the use of game design elements in non-game contexts," emphasizing its potential to trigger intrinsic motivation through elements such as points, levels, feedback, and challenges. In educational contexts, gamification has been shown to support persistence, autonomy, and competence-all of which are central components in self-determination theory [7]. Escape rooms embody gamified learning environments by transforming educational content into interactive challenges, engaging learners in dynamic and participatory tasks.

Hamari, Koivisto, and Sarsa [8] conducted a systematic literature review on gamification and found that game elements positively affect user engagement and motivation, although context-specific factors often mediate outcomes. In escape room settings, elements such as time constraints, feedback loops, and reward structures foster a sense of urgency and achievement, which can enhance both cognitive processing and emotional involvement.

These mechanics not only sustain engagement but also deepen learning by contextualizing abstract concepts in problem-solving scenarios.

Constructivist learning theory posits that learners actively construct knowledge through experiences, interactions, and reflection rather than passively receiving information. Rooted in the works of Piaget and Vygotsky, this theory emphasizes the role of social interaction and contextualized problem-solving in the learning process. Escape rooms provide an ideal constructivist learning environment by immersing participants in problem spaces where knowledge must be built collaboratively and iteratively [9]. Game-based learning, particularly when supported by narrative and role-play, allows for experiential and situated learning-core tenets of constructivism [10]. As students solve puzzles and navigate escape room challenges, they engage in authentic learning tasks that require hypothesis testing, critical thinking, and adaptive reasoning. These activities align with Vygotsky's [11] concept of the Zone of Proximal Development (ZPD), where learners perform tasks slightly beyond their individual capacity with the support of peers or facilitators, enhancing their learning potential.

Puzzle-solving has long been recognized as a medium for cognitive development. Problem-solving activities play a significant role in the development of operational thinking in children. Puzzles promote logical reasoning, spatial manipulation, pattern recognition, and sequential processing-all of which are critical for cognitive growth. They also support executive function skills such as working memory, cognitive flexibility, and planning [12].

In the context of escape rooms, puzzles are not merely recreational; they function as cognitive tools embedded within authentic scenarios. The real-time nature of escape room gameplay heightens mental engagement, as players must continuously assess, interpret, and act upon clues. Vygotsky [11] highlighted that such scaffolded challenges promote higher-order thinking by encouraging learners to integrate prior knowledge with situational cues. Moreover, the collaborative nature of escape room puzzles enhances metacognitive awareness, as players reflect on their strategies, successes, and failures throughout the process.

2.1 Classification of Puzzles in Learning Contexts

Puzzles used in escape rooms can be categorized into distinct cognitive domains, aligning with Gardner's theory of multiple intelligences [13]. Each puzzle type targets specific mental faculties, thereby supporting differentiated learning strategies.

Verbal-Linguistic Puzzles: These include crosswords, riddles, and word associations. They require semantic processing, vocabulary knowledge, and verbal reasoning. Such puzzles are especially effective in language acquisition and literacy development, engaging learners in lexical analysis and syntactic decoding [5, 13].

Logical-Numerical Puzzles: Logic grids, Sudoku, and number sequences fall under this category. They are designed to stimulate deductive reasoning, pattern recognition, and mathematical reasoning. These puzzles support abstract thinking and are particularly beneficial in subjects like mathematics, physics, and computer science [13, 14].

Visual-Spatial Puzzles: These involve image-based tasks such as spot-the-difference games, jigsaw puzzles, and symbol decoding. Visual-spatial puzzles enhance learners' ability to interpret visual information, recognize patterns, and navigate spatial relationships. They are useful in fields requiring visual literacy such as geography, architecture, and the arts [13].

Kinesthetic Puzzles: These require physical interaction with objects, including assembling parts, manipulating locks, or navigating physical mazes. Kinesthetic puzzles engage tactile learners and promote procedural memory. In escape rooms, they may include tasks like physically unlocking a box using a code derived from other puzzles or aligning physical objects in a specific sequence to trigger a mechanism [14].

These categories not only illustrate the versatility of puzzle formats but also affirm their alignment with educational goals across disciplines. When intentionally integrated, puzzle typologies in escape rooms provide a rich platform for multi-sensory, interdisciplinary learning experiences.

3. Methodology

This study adopts a conceptual methodology aimed at constructing a typology of puzzles used in escape rooms and analyzing their pedagogical functions within gamified learning environments. Rather than employing empirical data collection, this approach synthesizes findings from existing scholarly and gray literature to form a theoretical framework connecting puzzle typologies with cognitive, affective, and collaborative learning outcomes. The purpose of this methodology is to offer a structured and transferable model that educators and instructional designers can utilize when integrating escape room puzzles into formal and informal learning settings.

3.1 Puzzle Categorization Framework

The classification model developed in this paper is grounded in cognitive and educational psychology, particularly theories of multiple intelligences [13], constructivist learning [11], and gamification [6]. Puzzle types are categorized along two principal axes:

- Cognitive Modality: The type of mental or physical process required to solve the puzzle.
- Game Functionality: The role the puzzle plays in the overall escape room experience (e.g., linear progression, meta-clue synthesis, skill development) [5].

Based on this dual-axis model, four primary puzzle categories are established [3, 4, 14]:

- Verbal-Linguistic Puzzles: Crosswords, riddles, and word games that rely on language comprehension and semantic reasoning.
- Logical-Numerical Puzzles: Sequence problems, code-breaking, and mathematical grids like Sudoku, which activate logical deduction and numerical pattern recognition.
- Visual-Spatial Puzzles: Hidden object games, map navigation, and symbolic decoding, engaging spatial awareness and pattern recognition.
- Kinesthetic Puzzles: Tasks involving object manipulation, tactile clues, or bodily movement (e.g., rearranging physical items, unlocking boxes), which emphasize procedural and physical engagement.

Each type is examined not in isolation but as part of a puzzle ecology-a sequential and thematically integrated system in which puzzles interact to form a holistic learning experience. This ecological approach also accounts for meta-puzzles, where previously solved puzzles contribute to a final, overarching solution. Meta-puzzles are positioned as higher-order integrators within the game structure and are often aligned with synthesis-level cognitive tasks from Bloom’s taxonomy [15].

3.2 Mapping to Educational Outcomes

To establish the educational relevance of each puzzle type, this framework correlates puzzle modalities with measurable learning outcomes, drawing from established pedagogical taxonomies. The mapping is given in Table 1.

Table 1. Classification of puzzle types based on cognitive skills and associated educational outcomes.

Puzzle type	Primary cognitive skill	Associated educational outcome
Verbal-Linguistic	Semantic processing	Language acquisition, critical reading, vocabulary expansion
Logical-Numerical	Deductive reasoning	Mathematical literacy, logical thinking
Visual-Spatial	Pattern recognition	Visual literacy, diagram interpretation
Kinesthetic	Procedural memory, coordination	Motor skills, experiential learning
Meta-Puzzles	Synthesis, strategy planning	Higher-order reasoning, metacognition

This alignment supports the construct validity of the typology in educational settings. For instance, verbal-linguistic puzzles can be used to teach foreign language vocabulary through narrative context, while logical-numerical puzzles are suited to reinforcing algorithmic thinking in programming courses. Moreover, the collaborative nature of many escape room puzzles encourages communication, teamwork, and leadership skills-key 21st-century competencies [16].

Each puzzle type also corresponds with specific game mechanics associated with gamification, such as progression systems (e.g., unlocking sequential puzzles), feedback loops (e.g., correct input triggers a new clue), and reward structures (e.g., discovering a key or storyline fragment). This integration is guided by gamification design principles, such as meaningful play, clear objectives, and intrinsic feedback [17]. By embedding educational content within these mechanics, puzzles transform passive learning into active exploration.

The methodology thus not only categorizes puzzle types but also situates them within a broader framework of game-based pedagogy. This allows for the intentional design of escape rooms that align with learning outcomes, cognitive modalities, and motivational systems-making them replicable and adaptable across disciplines.

4. Puzzle Typologies and Function in Escape Rooms

The design and function of puzzles in escape rooms are integral to the overall gameplay experience, narrative coherence, and educational efficacy. Puzzle typologies-mental, physical, and meta-serve distinct but often interrelated cognitive and experiential functions. Their effective implementation depends on careful attention to game mechanics such as sequencing, theming, and hint systems. This section categorizes and examines these puzzle types, illustrating their role within the gamified escape room environment.

4.1. Mental Puzzles

Mental puzzles are cognitively demanding tasks that involve reasoning, memory recall, language processing, or mathematical deduction. Common examples include code-breaking sequences, symbolic logic chains, riddles, number-letter ciphers, and linguistic decodings. These puzzles typically require players to infer solutions based on abstract representations or pattern identification.

For instance, a numeric padlock with a three-digit code might be unlocked by solving a riddle embedded within the room's narrative: "You see me rise, but never set. I brighten your day without a regret." The answer "sun" may correspond to a quantity indicated visually (e.g., the number of suns in the room's artwork), leading to a numeric input. In another case, a Caesar cipher might be used to encode a message that, when deciphered, provides a clue to the next step [18].

Designing mental puzzles requires ensuring thematic consistency-puzzles must feel like a natural part of the environment and story-and providing a layered hint system that avoids frustrating players. These puzzles often anchor the linear or non-linear progression of the game, and their difficulty should scale appropriately across the duration of the experience [18].

4.2. Physical Puzzles

Physical puzzles require interaction with tangible elements in the escape room space, engaging players through object manipulation, spatial reasoning, and sometimes motor coordination. They are designed to appeal to kinesthetic learners and provide tactile and immersive engagement. Examples include reassembling disassembled physical mechanisms, aligning movable objects to a specific configuration, or navigating physical challenges such as crawling through a laser grid without triggering sensors.

One common design is a locked box that can be opened only when three statues in a room are rotated to face a particular direction, based on clues embedded in wall paintings. Another example is the classic "magnetic maze," where players use a magnet under a table to guide a metallic ball through a labyrinth, revealing a code or key at the end [19].

These puzzles depend heavily on affordances-i.e., how intuitive it is for players to know what can be touched, moved, or interacted with. Effective physical puzzles integrate seamlessly with the room's aesthetic design, reinforcing narrative immersion [19]. Importantly, physical puzzles must also be robust, given repeated physical use, and include resettable mechanisms for scalability in educational or commercial use.

4.3. Meta Puzzles

Meta puzzles function as overarching challenges that require the integration of multiple prior puzzle solutions. They are typically positioned near the conclusion of the escape room experience and serve as synthesis tasks, demanding that players abstract and combine knowledge gained throughout the game. Meta puzzles are often non-obvious in nature and involve what Johnson [20] terms "correlation logic"-the linking of seemingly disparate elements into a cohesive solution.

A representative meta puzzle might involve players collecting fragmented map pieces throughout the room, each obtained by solving smaller puzzles. Only once all fragments are acquired and assembled does the final exit route become clear. Another example is a multi-step logic puzzle in which players receive alphanumeric values from previous tasks, which must then be input in a specific order to trigger the final lock.

These puzzles exemplify Bloom's synthesis level, requiring players to recognize connections across knowledge domains and apply them strategically [15]. Their design must carefully balance difficulty and clarity, often using redundant clue structures to ensure solvability without compromising challenge. When successful, meta puzzles provide a climax to the narrative and gameplay arc, yielding a profound sense of achievement and closure.

4.4. Design Considerations Across Puzzle Types

Across all puzzle categories, three fundamental design elements are crucial for educational and experiential coherence:

- Hint Systems: Hints should be tiered, offering escalating levels of guidance without directly revealing answers. Automated hint delivery (e.g., time-based or triggered by incorrect input) can preserve immersion and reduce facilitator intervention.
- Sequential Flow: Puzzle progression must be either linear-requiring puzzles to be solved in a specific order-or modular, allowing simultaneous task engagement. The structure chosen affects team dynamics and learning opportunities [5].
- Thematic Consistency: All puzzles should support the narrative arc of the escape room. Incongruous or overly abstract puzzles may break immersion and reduce the educational impact. For example, in a forensic science-themed room, puzzles based on DNA sequences or fingerprint matching reinforce subject-specific content, aligning gameplay with curricular objectives [4].

When well-implemented, the integration of mental, physical, and meta puzzles within a coherent game loop not only sustains engagement but also enables meaningful learning experiences that blend cognitive challenge, collaboration, and narrative immersion.

5. Educational and Institutional Applications

Escape rooms have evolved from recreational pastimes into potent pedagogical and training tools across diverse educational and institutional settings. By embedding curriculum-aligned puzzles within immersive game environments, escape rooms facilitate experiential learning, formative assessment, and skills development. Their applications span formal education, corporate training, and professional development contexts, offering adaptable frameworks that can align with instructional goals across disciplines and industries [4].

5.1. Application in Classrooms

The integration of escape rooms into classroom instruction has shown strong potential to enhance learner engagement and deepen content understanding, particularly in subjects such as science, mathematics, and language arts. Studies have demonstrated that escape room activities can increase student motivation, activate prior knowledge, and encourage collaborative problem-solving [4]. For example, in science classrooms, puzzles involving DNA sequences, periodic tables, or chemical bonding can be designed as challenges requiring application of theoretical knowledge. In mathematics, logic-based puzzles such as Sudoku, coordinate mapping, or sequence identification can reinforce concepts like spatial reasoning, algebra, or arithmetic.

Classroom escape rooms can also be employed as review tools to consolidate learning before examinations or at the conclusion of instructional units. By transforming abstract or rote content into concrete and engaging tasks, these rooms promote long-term retention and provide real-time formative feedback. Furthermore, the social component of escape room collaboration enhances classroom cohesion and supports peer learning models aligned with constructivist pedagogy [10, 11].

5.2. Use in Corporate and Professional Training

In corporate and institutional contexts, escape rooms have been adapted as training simulations to develop competencies in communication, teamwork, leadership, and decision-making under pressure. The format is particularly effective for onboarding programs, team-building retreats, and executive development, where soft skills are critical. Unlike traditional workshops, escape rooms simulate real-world constraints-such as time pressure and information asymmetry-that compel participants to collaborate strategically, distribute roles, and adapt rapidly to emerging challenges [5].

For example, a cybersecurity firm may employ a digital escape room simulating a data breach scenario, requiring participants to decode firewalls, trace threats, and apply company protocols. Healthcare organizations have similarly utilized medical-themed escape rooms to reinforce diagnostic reasoning, emergency response coordination, and ethical decision-making, fostering critical reflection through experiential learning [21].

Escape rooms also support interdisciplinary learning and professional upskilling by offering gamified environments where domain knowledge, procedural fluency, and interpersonal skills intersect. These experiences can be customized to specific industry standards, regulatory frameworks, or organizational values, making them a versatile and scalable tool for human resource development.

5.3. Integration with Formal Assessments

Beyond training and review, escape rooms are increasingly employed as alternative assessment environments. They allow instructors to evaluate not only content mastery but also process skills such as collaboration, creativity, and resilience. The puzzle-solving nature of escape rooms aligns well with performance-based assessments, where students demonstrate learning through application rather than memorization [22].

Puzzle formats commonly used in escape rooms-such as matching terms to definitions, sequencing cause-effect relationships, identifying errors, or solving multiple-choice problems embedded in riddles-can be directly aligned with curriculum standards. These formats support a wide range of cognitive levels, from recall to synthesis, and offer a medium for differentiated assessment that accounts for diverse learner strengths.

For example, a history-themed escape room might include a chronological sequencing puzzle requiring students to order major events leading up to a conflict. A biology escape room might include a matching puzzle associating organ systems with their functions, while an economics escape room might feature fill-in-the-blank puzzles interpreting market data. These task types correspond to familiar question structures used in traditional exams, yet their contextualization in game scenarios reduces test anxiety and increases engagement [22].

Additionally, digital escape rooms can incorporate automated data tracking, enabling instructors to gather analytics on problem-solving approaches, time management, and teamwork dynamics. This opens new possibilities for learning analytics-driven assessment and personalized feedback.

5.4. Puzzle Formats Suitable for Pedagogy

Specific puzzle formats are especially well-suited for educational adaptation due to their clarity, flexibility, and ease of assessment. These include [4, 22]:

- Matching: Terms with definitions, causes with effects, formulas with outcomes.
- Multiple Choice: Embedded in riddles or as part of a locked mechanism requiring the correct option.
- Fill-in-the-Blanks: Often paired with visual or narrative clues; useful for language and content recall.
- Cause-Effect Chains: Chronological or logical progression puzzles, ideal for history, science, or systems-based learning.
- Error Detection: Identifying inaccuracies in a narrative or diagram, promoting critical reading and analysis.

These formats can be delivered through both analog and digital means and can be embedded in either linear or modular puzzle sequences. Their integration within escape room design supports both active learning and formative assessment, aligning well with 21st-century pedagogical priorities.

6. Technological Integration

As escape rooms evolve beyond their physical constraints, digital technologies such as Augmented Reality (AR), Virtual Reality (VR), and Internet of Things (IoT) devices are reshaping puzzle design, interactivity, and scalability. These innovations not only expand the possibilities for immersion and narrative complexity but also increase accessibility and pedagogical relevance in formal and informal learning environments. However, the transition from analog to digital also introduces technical and instructional design challenges that must be critically evaluated [23, 24].

6.1. Augmented Reality (AR) in Puzzle Interaction

AR enables the layering of digital elements-such as images, text, and animations-over physical environments through devices like smartphones, tablets, or smart glasses. In escape room contexts, AR transforms static clues into dynamic, interactive content that can adapt based on user input or real-time environmental cues. For example, players might point a tablet camera at a bookshelf to reveal an invisible code overlaid digitally, or scan a QR marker to trigger a character's holographic message offering the next clue [23].

This technology enhances spatial and contextual learning by integrating multimodal stimuli into puzzle interaction. It is particularly valuable in educational settings where AR-based puzzles can simulate scientific phenomena, historical reenactments, or language translation scenarios, allowing for authentic task engagement [23]. Moreover, AR reduces reliance on physical props, enabling modular and reusable puzzle components in hybrid learning spaces.

Nevertheless, AR implementation requires a robust infrastructure, including stable internet connections, compatible hardware, and reliable software frameworks. User experience can be diminished by poor tracking accuracy, device incompatibility, or cognitive overload caused by excessive visual elements [24].

6.2. Virtual Reality (VR) for Immersive Puzzle-Based Storytelling

VR provides a fully immersive digital environment in which players can engage with puzzles and narratives beyond the limitations of physical space. Using head-mounted displays and motion controllers, learners can explore simulated environments, manipulate virtual objects, and interact with computer-generated avatars. In escape room scenarios, VR can simulate settings impractical in real life-such as deep-sea labs, ancient ruins, or space stations-thus broadening the thematic and instructional palette [25].

VR is particularly effective in promoting presence, a psychological state in which users feel physically and emotionally located within the virtual environment [26]. This presence amplifies cognitive engagement and can support learning in complex domains such as engineering, medicine, or environmental science. For instance, a chemistry VR escape room might require players to combine elements using a virtual periodic table to simulate a safe chemical reaction, reinforcing conceptual understanding through embodied interaction [25].

However, the immersive nature of VR also presents pedagogical and technical challenges. These include motion sickness, hardware cost, limited access, and the need for highly skilled instructional designers familiar with 3D modeling and game development platforms such as Unity or Unreal Engine [25]. Furthermore, VR experiences must be carefully scaffolded to prevent cognitive overload and ensure that technological novelty does not overshadow learning objectives.

6.3. Robotics and IoT-Based Puzzles

Robotic elements and IoT devices enable the integration of responsive, programmable components into escape room environments. These might include electronic locks, sensors, motors, or voice-activated devices that respond to player actions. For example, solving a numeric puzzle on a touchscreen might trigger a servo motor to open a hidden compartment, or completing a sound-based task might cause a robot to deliver the next clue [27].

Such technology supports procedural learning, where students interact with systems that mimic real-world automation and control logic. This is particularly relevant in STEM education, where learners can experience embedded systems, coding, and mechanical feedback loops in a hands-on context [27]. In professional training, IoT-based puzzles can simulate operational environments, such as manufacturing lines or smart building systems.

Despite these benefits, integrating robotics and IoT components requires significant planning, from circuit design and power management to firmware programming. Maintenance and reliability can become issues in high-use environments, especially where precise calibration is needed. Additionally, safety considerations must be addressed when physical machinery interacts with users.

6.4. Advantages and Challenges of Digitalizing Escape Room Experiences

The digitalization of escape rooms offers several pedagogical and logistical advantages. Digitally enhanced puzzles are more easily scalable, modifiable, and distributable across physical or virtual locations. They allow for individualized learning paths, automatic data collection for assessment, and integration with Learning Management Systems (LMS). Moreover, digital platforms facilitate asynchronous learning, enabling remote and hybrid escape room experiences that accommodate diverse learner needs [22].

However, the transition to digital formats is not without trade-offs. Key challenges include [22, 25]:

- Loss of tactile engagement: Kinesthetic learners may struggle in fully digital environments that lack physical manipulation.
- Technical complexity: The design and deployment of AR/VR/IoT-enhanced puzzles demand interdisciplinary expertise.
- Equity and access: Not all learners or institutions possess the necessary hardware or bandwidth to participate fully.
- Design fatigue: Overuse or poorly integrated technology can lead to disengagement or distraction from learning goals.

Ultimately, successful technological integration in escape room puzzles requires a balanced approach-one that enhances interactivity and learning outcomes without overwhelming users or compromising accessibility.

7. Design Considerations and Best Practices

Effective escape room design requires more than assembling engaging puzzles; it demands intentional planning to ensure pedagogical alignment, narrative coherence, and inclusive accessibility. Well-crafted escape rooms harmonize game mechanics with learning outcomes, emotional engagement, and logistical feasibility. This section presents best practices for escape room puzzle design by focusing on four interrelated pillars: puzzle clarity and narrative relevance, logical progression, fail-safe operations, and inclusive access.

7.1. Puzzle Clarity and Relevance to Story

One of the most critical elements of escape room design is the semantic and contextual clarity of puzzles. Puzzles must be intelligible within the constraints of the room's theme and understandable to players with varying levels of prior knowledge. Ambiguity in language, unclear objectives, or confusing clues can disrupt player immersion and learning flow. Narrative integration enhances both engagement and cognitive alignment. Each puzzle should serve a story-driven function-advancing the plot, revealing character backstories, or simulating a critical moment in the scenario. For example, in a forensic-themed room, using fingerprint analysis as a puzzle not only reinforces scientific content but also sustains thematic immersion [5].

To maintain clarity, designers are encouraged to follow the "one puzzle, one purpose" principle-ensuring each puzzle has a specific role and avoids redundant or superfluous complexity. Instructions should be embedded in the environment through intuitive design elements such as color coding, spatial alignment, or symbolic cues rather than extensive written directions [5].

7.2. Internal Consistency and Logic in Puzzle Progression

Escape rooms must exhibit internal logical coherence, where the sequence of puzzles follows a rational, consistent structure. Players should be able to anticipate, infer, and build upon previous solutions in a way that reflects problem-solving logic. Inconsistent mechanics, unpredictable difficulty spikes, or disconnected puzzles can lead to player frustration and disengagement.

A well-designed escape room operates through either linear progression (where each puzzle unlocks the next in a strict order) or modular/non-linear progression (where puzzles can be solved in parallel but converge at a meta-puzzle or final challenge). Both models require clear signposting, i.e., visual or structural indicators that guide players through the experience without explicit instruction [3].

Moreover, each puzzle must contribute to a difficulty curve that balances challenge and reward. Cognitive load theory suggests that optimal puzzle progression should increase complexity gradually, allowing learners to consolidate foundational knowledge before tackling higher-order tasks [28].

7.3. Testing and Contingency Planning

Rigorous playtesting is essential to validate the functionality, fairness, and flow of puzzles. Every element-mechanical or digital-should be tested for usability, interpretability, and durability under repeated use. Common issues such as unclear clues, ambiguous solutions, or technical malfunctions can significantly undermine the learning experience [22].

Designers must also account for contingency scenarios, including mechanical failures, software crashes, or participant confusion. Backup plans may include [22]:

- Duplicate copies of essential physical components (e.g., keys, props).
- Manual overrides or hint triggers controlled by a facilitator.
- Digital redundancies, such as parallel input methods (e.g., keypad and touchscreen).
- Scripted interventions to maintain immersion while addressing breakdowns.

In educational contexts, where time constraints and learning objectives are non-negotiable, contingency planning ensures that the escape room remains functional and aligned with instructional goals regardless of unforeseen disruptions.

7.4. Accessibility and Inclusivity in Puzzle Design

Designing for accessibility and inclusivity ensures that all participants, regardless of physical ability, cognitive style, or cultural background, can meaningfully engage with escape room activities. Inclusive design not only adheres to universal design principles but also enhances the pedagogical value by accommodating diverse learner needs [29].

Physical accessibility considerations include ensuring that puzzles do not rely on dexterity, fine motor skills, or mobility beyond what can be reasonably accommodated. For example, puzzles that require reaching high shelves or navigating narrow spaces should have alternative solutions or adjustments for participants with mobility impairments.

Cognitive accessibility involves minimizing unnecessary complexity, avoiding culturally biased references, and ensuring clarity in instructions and expected actions. For example, riddles or wordplay that depend on idiomatic language may disadvantage non-native speakers and should be used judiciously or accompanied by visual scaffolds [22, 29].

Audio and visual elements should be multi-modal, allowing players to engage through alternative sensory channels. Providing written text along with sound cues, or tactile interfaces in addition to visual puzzles, can broaden accessibility [24]. Additionally, diversity and representation in character design, narratives, and scenarios help foster an inclusive environment that reflects a variety of identities and experiences.

8. Discussion

Escape rooms represent a convergence of game design, pedagogy, and experiential learning, with puzzle typologies acting as the primary mediators between engagement and educational value. As demonstrated across this paper, distinct puzzle categories-mental, physical, and meta-correspond to specific cognitive domains and foster a wide range of transferable skill sets [4, 22]. These puzzles, when intentionally designed and pedagogically aligned, serve not only to challenge but to transform how learners interact with content, collaborate with peers, and reflect on problem-solving processes.

Each puzzle type cultivates a unique constellation of skills. Mental puzzles, such as logic chains and riddles, promote abstract reasoning, verbal fluency, and deductive logic. These tasks activate analytical thinking and require players to apply known principles to novel situations-skills central to academic success across disciplines. Physical puzzles, by contrast, engage procedural memory, spatial reasoning, and motor coordination. These tasks are invaluable in developing hands-on competencies, especially in STEM and vocational fields where physical manipulation of tools and systems is essential [27]. Meta puzzles, which demand the integration of multiple puzzle solutions into a comprehensive final task, foster high-level synthesis, strategic planning, and systems thinking, mirroring the cognitive demands of real-world problem-solving [15].

Beyond cognitive skills, all puzzle types encourage metacognitive reflection and soft skill development, including perseverance, adaptability, and collaborative communication. The social architecture of escape rooms necessitates that players share information, negotiate roles, and resolve conflicts in real-time-behaviors aligned with teamwork and leadership competencies essential in both academic and professional environments [22].

8.1. Escape Rooms as Transdisciplinary Tools

Escape rooms are inherently transdisciplinary, functioning across domains such as education, entertainment, corporate training, and therapeutic intervention. In education, their adaptability to curricular content makes them suitable for virtually any subject area, from history and biology to computer science and foreign languages. Their integration with assessment practices, differentiated instruction, and inclusive design enables application in primary, secondary, and tertiary contexts [4, 5].

In corporate and professional development settings, escape rooms serve as experiential simulations that replicate organizational dynamics under controlled stress conditions. These experiences are particularly effective in evaluating team cohesion, communication efficiency, and ethical decision-making. Similarly, escape rooms are increasingly explored in therapeutic contexts-especially in cognitive-behavioral therapy, trauma treatment, and occupational therapy-as controlled environments for practicing emotional regulation, social interaction, and motor functions [30].

The capacity of escape rooms to operate across these diverse contexts lies in their foundational design: they are modular, immersive, and outcome-oriented. When framed appropriately, puzzles become vessels for learning, self-discovery, and even rehabilitation-demonstrating that gamified experiences are not confined to entertainment but possess transformative potential across domains [29].

8.2. Implications for Future Educational Technologies and Gamified Platforms

As educational technology continues to evolve, escape rooms offer a compelling template for next-generation learning environments. Their reliance on narrative, challenge, and interactivity aligns with contemporary learning theories that emphasize learner agency, contextualization, and experiential depth. The integration of emerging technologies such as Augmented Reality (AR), Virtual Reality (VR), and Internet of Things (IoT) will further enhance the responsiveness, scalability, and personalization of escape room experiences [25].

Future gamified platforms inspired by escape room dynamics could serve as adaptive learning ecosystems, capable of real-time data tracking, performance analytics, and personalized scaffolding. These systems may incorporate AI-driven hints, dynamically adjusting difficulty levels, or offer alternate narrative paths based on learner choices, creating a hybrid of escape room gameplay and intelligent tutoring systems.

However, realizing this vision requires interdisciplinary collaboration between educators, game designers, psychologists, and technologists. Challenges such as digital equity, cognitive overload, and sustainability must be addressed to ensure that technological enhancement does not overshadow educational intent. Additionally, research must continue to explore the longitudinal effects of escape room learning on knowledge retention, skill transfer, and learner motivation [22].

Ultimately, the escape room model embodies the core principles of 21st-century education: active learning, collaboration, problem-solving, and creativity. By continuing to evolve in tandem with educational technologies, escape rooms have the potential to become foundational structures in future learning paradigms.

9. Conclusion

This conceptual study has outlined a comprehensive framework for understanding escape room puzzles as multidimensional tools within gamified learning environments. By categorizing puzzles into mental, physical, and meta types, the paper demonstrated how each puzzle modality supports distinct cognitive and affective skill sets, including problem-solving, collaboration, spatial reasoning, and strategic thinking. These puzzle typologies not only guide the design of engaging escape room experiences but also offer a pedagogical structure for aligning game mechanics with learning outcomes.

The analysis has further emphasized the educational potential of escape rooms in both formal and informal contexts. From classroom instruction in STEM and humanities disciplines to corporate training and therapeutic interventions, escape rooms have proven adaptable to diverse learning needs. Their ability to integrate content mastery with 21st-century skills-such as creativity, teamwork, and metacognition-positions them as valuable assets in instructional design and competency-based education. As reviewed, escape rooms can function as assessment tools, motivation enhancers, and collaborative platforms that facilitate experiential and student-centered learning.

Moreover, the study has mapped current technological trends transforming escape rooms into scalable, interactive, and immersive learning environments. The integration of Augmented Reality (AR), Virtual Reality (VR), and IoT-based puzzle mechanisms enables new levels of personalization, accessibility, and engagement. While these innovations introduce new complexities in design and deployment, they also expand the applicability of escape rooms in hybrid and remote learning models, aligning well with global shifts toward digital education infrastructure.

Taken together, these contributions underscore the growing relevance of escape games as lifelong learning tools. In a world increasingly shaped by rapid technological change, information overload, and evolving workplace demands, escape rooms offer a flexible and engaging medium through which learners of all ages can develop essential cognitive, emotional, and social skills. Their narrative-driven, problem-based structure mirrors the complexities of real-world challenges, making them not only a tool for academic development but also a framework for personal and professional growth.

As educational systems seek to foster deeper engagement and adaptability in learners, escape rooms-and the puzzle-based learning they promote-stand out as a pedagogical innovation with enduring relevance. Continued interdisciplinary research, combined with practitioner-driven experimentation, will be key to harnessing the full potential of escape rooms in shaping future-ready learners across all stages of life.

References

[1] Masacı D. Kaçış oyunları ve bulmacalara oyunlaştırma ile bakış. 2023. Retrieved from https://www.toligames.com/blog/icerik/bulmaca-nedir

[2] Sarıkaya M. Türkiye Cumhuriyeti'nin İlk Yıllarında Yayınlanan Çocuk Dergilerindeki "Akıl Oyunları" Örnekleri. Munzur Üniversitesi Sosyal Bilimler Dergisi. 2018;7(13):10-21.

[3] Wiemker M, Elumir E, Clare A. Escape room design theory. Escape Room Games: "Can you transform an unpleasant situation into a pleasant one?". 2015. Retrieved from https://thecodex.ca/wp-content/uploads/2016/08/00511Wiemker-et-al-Paper-Escape-Room-Games.pdf

[4] Borrego C, Fernández C, Blanes I, Robles S. Room escape at class: Escape games activities to facilitate the motivation and learning in computer science. Journal of Technology and Science Education. 2017;7(2):162-171.

[5] Nicholson S. Peeking behind the locked door: A survey of escape room facilities. White Paper. 2015. Retrieved from https://scottnicholson.com/pubs/erfacwhite.pdf

[6] Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: Defining "gamification". Proceedings of the 15th international academic MindTrek conference. 2011:9-15.

[7] Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist. 2000;55(1):68-78.

[8] Hamari J, Koivisto J, Sarsa H. Does gamification work? A literature review of empirical studies on gamification. 2014 47th Hawaii international conference on system sciences. 2014:3025-3034.

[9] McIlwhan R. Previous next view larger image educational escape rooms: exploring the potential of gamification in learning. 2024. Retrieved from https://lta.hw.ac.uk/educational-escape-rooms-exploring-the-potential-of-gamification-in-learning/

[10] Gee JP. What video games have to teach us about learning and literacy. New York: Palgrave Macmillan, 2007.

[11] Vygotsky LS. Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press, 1978.

[12] HOUDINI'S. Feel the adrenaline rush. 2024. Retrieved from https://www.houdinisescape.co.uk/blog/the-psychology-of-escape-rooms/

[13] Gardner H. Frames of mind: The theory of multiple intelligences. New York, NY: Basic Books, 1983.

[14] ERM (Escape Room Melbourne). 25 Types of Escape Room Puzzles. 2025. Retrieved from https://escaperoom.com.au/types-of-escape-room-puzzles/

[15] Bloom BS. Taxonomy of educational objectives: The classification of educational goals. New York: Longmans, Green, 1956.

[16] Trilling B, Fadel C. 21st century skills: Learning for life in our times. San Francisco, CA: Jossey-Bass, 2009.

[17] Kapp KM. The gamification of learning and instruction: Game-based methods and strategies for training and education. San Francisco, CA: Pfeiffer, 2012.

[18] Clare A. Escape room design theory. Escape Room Games: "Can you transform an unpleasant situation into a pleasant one?". 2015. Retrieved from https://thecodex.ca/wp-content/uploads/2016/08/00511Wiemker-et-al-Paper-Escape-Room-Games.pdf

[19] Sauerland V. Horror in escape rooms: How to build an atmosphere and design puzzles, Bachelor's thesis, Theseus.fi, Finland, 2023.

[20] Johnson C. Questions asked by Martin Mathers from Paragon Publishing. 2003. Retrieved from https://web.archive.org/web/20050310230359/http://www.thefoolsgold.com/CJ/2nd-interview.htm

[21] Backlund P, Hendrix M. Educational games – Are they worth the effort? A literature survey of the effectiveness of serious games. Proceedings of the 5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES). 2013:1-8.

[22] Veldkamp A, van de Grint L, Knippels MCPJ, van Joolingen WR. Escape education: A systematic review on escape rooms in education. Educational Research Review. 2020;31:100364.

[23] Billinghurst M, Clark A, Lee G. A survey of augmented reality. Foundations and Trends in Human–Computer Interaction. 2015;8(2-3):73-272.

[24] Akçayır M, Akçayır G. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. Educational Research Review. 2017;20:1-11.

[25] Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Computers & Education. 2020;147:103778.

[26] Slater M, Wilbur S. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. Presence: Teleoperators & Virtual Environments. 1997;6(6):603-616.

[27] Lathwesen C, Belova N. Escape rooms in STEM teaching and learning-Prospective field or declining trend? A literature review. Education Sciences. 2021;11(6):308.

[28] Sweller J. Cognitive load during problem solving: Effects on learning. Cognitive Science. 1988;12(2):257-285.

[29] Flynn CO. Escape to Learn: Digital Escape Rooms and Universal Design for Learning. All Ireland Journal of Teaching and Learning in Higher Education (AISHE-J). 2024;16(2).

[30] Brown S, Vaughan C. Play: How it shapes the brain, opens the imagination, and invigorates the soul. New York: Avery, 2018.

Research & Design
www.jrede.org

Review Article

A comparative review of hallucination mitigation and performance improvement techniques in Small Language Models

Fatih Ahmet Senel[1a], Hayri Baytan Ozmen*[1b]

[1]*Faculty of Engineering and Natural Sciences, Süleyman Demirel University, Isparta, Türkiye.*
[2]*Faculty of Engineering and Natural Sciences, Uşak University, Uşak, Türkiye.*

Article Info

Article History:

Received:
23 May 2025

Accepted:
15 June 2025

Keywords:

Knowledge Distillation;

Natural Language Processing (NLP);

Parameter-Efficient Fine-Tuning;

Performance optimization;

Retrieval-Augmented Generation (RAG);

Text generation

Abstract

Small Language Models (SLMs) offer a computationally efficient alternative to Large Language Models (LLMs), enabling natural language processing (NLP) capabilities in resource-constrained and/or private environments such as personal computers, mobile devices, embedded systems, and real-time applications. However, SLMs face significant challenges related to factual hallucination, limited generalization, and degraded task performance due to their reduced parameter capacity. This review provides a comparative analysis of current methods to mitigate hallucination and enhance performance in SLMs. We consider hallucination prevention techniques into five primary strategies: retrieval-augmented generation (RAG), instruction tuning and prompt engineering, fact-checking and verification layers, calibration mechanisms, and fine-tuning with human feedback. In parallel, we explore performance enhancement methods including quantization, pruning, parameter-efficient tuning, knowledge distillation, mixture-of-experts architectures, and domain-adaptive training. A comparative evaluation highlights trade-offs between accuracy, compute efficiency, and deployment feasibility. We identify best-fit combinations of techniques for diverse real-world scenarios—ranging from mobile applications to safety-critical systems—and discuss integration challenges, sustainability concerns, and ethical implications. This work is aimed to introduce and compare methods for developing robust and efficient SLMs capable of reliable deployment across varied NLP contexts.

1. Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence concerned with the interaction between computers and human (natural) languages. Its development has spanned several decades, from rule-based symbolic approaches in the 1950s and 1960s to statistical models in the 1990s and, more recently, neural models that dominate the field today [1, 2]. Early systems relied heavily on handcrafted grammars and lexicons, which limited their scalability and adaptability. With the advent of machine learning, particularly supervised learning, NLP began leveraging large annotated corpora to learn statistical patterns for tasks such as part-of-speech tagging, named entity recognition, and machine translation.

The transition to deep learning accompanied in the era of neural language models. Word embeddings such as Word2Vec [3] and GloVe [4] provided dense vector representations of

*Corresponding author: baytan.ozmen@usak.edu.tr
[a]orcid.org/0000-0003-1918-7277; [b]orcid.org/0000-0001-6750-8632;
DOI: http://dx.doi.org/10.17515/rede2025-004en0523rs

words, capturing semantic similarities. These were later surpassed by contextual embeddings produced by deep neural networks such as ELMo [5], BERT [6], and GPT [7, 8]. These models, especially the Transformer-based architectures, revolutionized NLP by enabling pretraining on massive text corpora followed by fine-tuning on specific downstream tasks.

Large Language Models (LLMs) such as GPT-3 [9], PaLM [10] and GPT-4 [11] possess billions of parameters and exhibit emergent capabilities such as few-shot learning, reasoning, and dialogue generation. These models have become foundational in a variety of applications, including virtual assistants, summarization tools, and generative text engines. However, LLMs also face significant challenges. Main among them are computational inefficiency, high energy costs, and difficulties in deployment on edge devices [12]. Additionally, LLMs are known to hallucinate—i.e., generate fluent but factually incorrect or misleading outputs—which poses serious risks in real-world applications [13].

In response to the operational challenges posed by LLMs, the NLP community has seen renewed interest in Small Language Models (SLMs)—models with significantly fewer parameters that are tailored for efficient, cost-effective deployment. SLMs are often used in scenarios requiring low-latency inference, privacy preservation, or deployment on devices with limited computational resources such as personal computers, mobile phones or embedded systems [14]. Although smaller in size, SLMs can still achieve competitive performance on specific tasks when equipped with architectural and training optimizations [15].

However, SLMs are more prone to issues such as hallucination, reduced reasoning capacity, and brittleness across domains due to their limited parameter space. Addressing these issues is critical for expanding their utility and reliability in real-world applications.

This review aims to provide a comprehensive overview of existing methods to prevent hallucination and enhance the performance of SLMs in NLP. Specifically, we:

- Introduce and explain key hallucination mitigation techniques such as retrieval-augmented generation (RAG), instruction tuning, and post-hoc verification;
- Examine performance enhancement strategies including parameter-efficient tuning (e.g., LoRA), quantization, pruning, and knowledge distillation;
- Offer a detailed comparative analysis of these techniques based on their efficacy, computational cost, and suitability for different use cases;
- Discuss when and how combinations of methods may be employed synergistically;
- Address the question of whether the performance gains justify the additional resource overhead, particularly in low-resource environments.

By focusing on SLMs, this review tries to fills a gap in the literature where most reviews tend to prioritize LLMs. We aim to equip researchers and practitioners with the insights needed to select and implement optimal strategies for building accurate, efficient, and responsible small language models.

2. Understanding Hallucination in Language Models

In the context of natural language generation (NLG), hallucination refers to the generation of output that is linguistically plausible but factually inaccurate, unverifiable, or misleading. Although hallucination affects models of all scales, its manifestations and impacts are often amplified in Small Language Models (SLMs) due to their reduced capacity and generalization ability [13].

The taxonomy of hallucination can be divided broadly into two categories [16, 17]:

- Intrinsic Hallucinations: These are errors introduced due to model misinterpretation or lack of understanding of the input context. The generated content contradicts or deviates from the input or ground truth. For example, in abstractive summarization, an intrinsic hallucination might involve the model fabricating an event that is not present in the source document.
- Extrinsic Hallucinations: In these cases, the model generates information that is not explicitly supported or contradicted by the input. These statements may seem plausible but cannot be grounded in the provided data. This form of hallucination is particularly common when the model is prompted with underspecified queries or lacks external factual support.

More recent studies have extended this classification to include fabricated references, nonsensical reasoning, and domain transfer hallucinations, especially in complex tasks such as dialogue generation, open-domain question answering, and summarization of long-form documents [18].

2.1 Causes of Hallucination in SLMs

While hallucination is a known issue even in large models like GPT-3 and PaLM, it is exacerbated in SLMs due to their constrained architecture, limited training data, and restricted contextual awareness[13]. Below are key contributing factors to hallucinations in SLMs:

2.1.1 Limited model capacity

SLMs typically operate with tens to hundreds of millions of parameters, compared to the tens of billions in LLMs. This limitation reduces their ability to represent complex semantic relationships, long-range dependencies, and nuanced contextual signals [14]. As a result, SLMs often generalize poorly, especially in tasks requiring factual precision or cross-sentence coherence.

2.1.2 Shallow contextualization and shorter attention windows

Many SLMs are trained or fine-tuned with limited input context (e.g., 512 tokens or fewer), which constrains their ability to understand extended discourse or refer to multiple evidential sources. In multi-turn conversations or document-level summarization, this results in hallucinations due to the loss of grounding signals across segments [19].

2.1.3 Training data biases and incompleteness

Hallucinations often stem from gaps in training corpora or exposure to unreliable text [20]. Since SLMs are trained on a subset of what LLMs typically ingest, they may lack exposure to edge cases, rare facts, or formal domain-specific knowledge, increasing the risk of error in critical use cases such as biomedical NLP or legal summarization.

2.1.4 Over-reliance on statistical priors

SLMs, like their larger counterparts, optimize the next-token prediction probability, leading them to generate text that is statistically probable rather than factually correct. This is particularly problematic when the model is asked for specific, non-frequent information that diverges from corpus norms, such as citing an uncommon journal article or describing niche technical concepts [21].

2.1.5 Lack of external grounding

Most SLMs do not include mechanisms for retrieval-augmented generation (RAG) or fact-checking modules that can verify their outputs against real-time knowledge sources. Without such grounding, these models are forced to rely solely on internal representations, increasing the likelihood of hallucinating when asked for factual information or summarizing ambiguous input [22].

2.1.6 Catastrophic forgetting in fine-tuning

When SLMs are fine-tuned for downstream tasks without care, they may "forget" previously learned knowledge due to the phenomenon of catastrophic forgetting [23]. This often results in hallucinations, especially when fine-tuning for domains with conflicting or newly introduced knowledge representations.

2.2 Impacts of Hallucination on Downstream Applications

The presence of hallucinations in SLM-generated content poses significant technical, ethical, and operational challenges, especially in applications where trust and accuracy are non-negotiable.

In fields like medicine and law, where AI tools are increasingly deployed to generate summaries, case reports, or recommendations, hallucinated content can result in diagnostic errors, misinterpretation of legal precedents, and regulatory non-compliance [24, 25]. SLMs deployed in low-resource clinical settings, if not robustly trained and evaluated, can propagate dangerous misinformation.

When used in education, hallucinations can mislead learners with inaccurate explanations or fabricated references, undermining trust in automated tutoring systems or language learning assistants. These errors are often subtle and hard to detect, especially in SLMs trained on general-purpose data.

For chatbots and digital assistants, hallucinations degrade user experience and erode trust. Users may take plausible-sounding but false statements at face value, particularly when the hallucinations are delivered fluently and confidently. In domains such as finance or insurance, this can have legal and reputational consequences [19].

Hallucinations contribute to content pollution when SLMs are used to generate or amplify content at scale, especially in social media or low-cost content farms. The ability of even small models to produce vast volumes of text means that low-quality or fabricated content can easily outpace human fact-checking and moderation efforts.

2.3 Hallucination Severity in Low-Resource vs. High-Resource Settings

Hallucinations manifest with greater severity and frequency in low-resource environments, which include both data-scarce tasks and deployment scenarios with limited compute, memory, or connectivity. In such cases, the model's ability to access external grounding sources or undergo thorough fine-tuning is severely constrained. For example, in healthcare delivery in rural areas, SLMs may be deployed on mobile devices with minimal local data and no internet access. Without retrieval augmentation or updated knowledge bases, these models rely entirely on internal parameters trained on outdated or general-purpose corpora, thereby increasing their tendency to hallucinate [19, 26].

In contrast, high-resource settings may integrate SLMs with supporting infrastructure like retrievers, validators, or human-in-the-loop supervision, allowing for multi-stage generation pipelines that mitigate hallucination. However, even in these settings, when computational priorities shift toward real-time response or user privacy, grounding mechanisms may be disabled, again increasing hallucination risk. This contrast suggests that deployment context significantly shapes the impact of hallucinations, and that a one-size-fits-all solution is unlikely to suffice. Fine-grained trade-off analyses between computational cost, hallucination risk, and application criticality are necessary.

2.4 Implications for Trust, Safety, and Model Evaluation

As hallucinations become a central concern in language model deployment, especially in SLMs intended for real-world use, the following implications must be considered:

Even when hallucinations are rare, the inability to predict when a hallucination will occur erodes user trust. Trust calibration requires that users can distinguish between high- and low-confidence model responses, ideally through probabilistic output scores, uncertainty estimations, or fact-checking signals [27]. Unfortunately, most SLMs do not natively provide such mechanisms and must rely on auxiliary components.

Current evaluation metrics like BLEU, ROUGE, or even BERTScore focus on fluency and lexical similarity, but fail to capture factual correctness [28]. Recent proposals such as FactCC [29], SummaC [30] and TruthfulQA [31] attempt to measure hallucination explicitly, but are primarily calibrated for large models. The lack of SLM-specific hallucination benchmarks limits both the diagnosis and mitigation of hallucination in these smaller models.

For regulated industries or safety-critical applications, hallucinations may not only be unacceptable but also legally actionable. As such, regulatory bodies and enterprises are increasingly requiring explainability, traceability, and auditable decision-making in language models, including SLMs. This necessitates hallucination prevention not just as a technical fix but as part of a broader model governance strategy [32].

3. Performance Challenges in Small Language Models

3.1 Memory and Computational Constraints

A defining characteristic of Small Language Models (SLMs) is their architectural minimalism—typically comprising under a few billion to 500 million parameters, sometimes fewer than 100 million. While this parameter reduction provides benefits in terms of deployment efficiency, energy consumption, and latency, it inherently limits the model's ability to represent, store, and generalize over complex language patterns [14].

The compact size of SLMs reduces the breadth of linguistic and world knowledge that can be encoded in model weights. This becomes particularly evident in complex tasks requiring multiple reasoning steps, deep factual recall, or cross-document synthesis. Moreover, many SLMs must operate under stringent inference constraints in environments such as smartphones, microcontrollers, and edge devices. These settings often restrict the available RAM, storage bandwidth, and energy budget, forcing additional compromises like model quantization or the use of limited-context token windows [33].

Crucially, these limitations also restrict the use of advanced processing techniques such as ensemble methods, retrieval-augmented generation, or large-scale context chaining—all of which are feasible with LLMs but impractical or infeasible in SLM regimes [34]. As a result, SLMs often suffer from underfitting, especially in long-form text generation and tasks requiring structured reasoning.

3.2 Limitations in Representation and Generalization

SLMs must learn to compress vast linguistic structures into compact parameter spaces, often resulting in weaker generalization performance on unseen or rare inputs. Unlike large models that possess significant capacity—allowing them to memorize and interpolate from large corpora—SLMs operate near the threshold of their learning ability, making them more susceptible to data sparsity, task shift, and domain drift [35].

This leads to several concrete limitations:

- Poor transfer learning: When pre-trained SLMs are adapted to new tasks or domains, they frequently exhibit catastrophic forgetting, where prior knowledge is overwritten by new information [23]. This is especially problematic in multi-domain deployments, where the same model must generalize across different styles, topics, or objectives.
- Reduced contextual sensitivity: Due to smaller attention heads and fewer layers, SLMs often fail to model long-range dependencies in text, which are critical for tasks such as summarization, multi-hop question answering, or document-level sentiment analysis [36].
- Increased reliance on statistical priors: Because of insufficient learning capacity, SLMs tend to default to high-frequency or "safe" outputs from the training distribution, limiting creativity and adaptability in dynamic or evolving environments [37].

In practical applications, these limitations surface as bland, repetitive, or overly generic outputs, particularly in tasks requiring open-ended generation or abstraction beyond surface-level inputs.

3.3 Trade-offs Between Efficiency and Accuracy

A central tension in the design of SLMs is the trade-off between computational efficiency and model accuracy. While SLMs are designed to be lightweight and fast, these benefits often come at the cost of reduced performance on standard NLP benchmarks—especially on tasks requiring semantic nuance, factual grounding, or task-specific adaptation [38].

Trade-off 1: Latency vs. Expressivity

Models designed for fast response—such as those used in interactive systems or edge inference—are optimized for minimal latency, which often necessitates shallow networks, aggressive quantization, or pruning. These operations reduce the model's expressive power and tend to impair output richness and contextual fidelity [39].

Trade-off 2: Generalization vs. Specialization

Highly compact SLMs trained for general use frequently underperform in domain-specific settings, such as legal, biomedical, or financial NLP [40]. Conversely, models fine-tuned for domain specificity may perform poorly on general language tasks [41]. Unlike LLMs, which can afford to retain multipurpose capacities, SLMs often face sharp trade-offs between domain specialization and task versatility [42, 43].

Trade-off 3: Interpretability vs. Optimization

Certain SLM configurations—such as those using low-rank approximations or distilled architectures—offer better interpretability due to their simpler structure, but may not perform optimally on complex tasks [14, 44]. This makes it challenging to strike a balance between transparency and task efficacy, particularly when explainability is a regulatory or business requirement.

Overall, while SLMs hold promise for scalable and democratized NLP, their design is fundamentally constrained by these trade-offs. The implications are profound: achieving performance gains in SLMs requires clever engineering, tailored training regimens, and often, external augmentation strategies—topics that will be discussed in subsequent sections.

4. Hallucination Prevention Techniques

The challenge of hallucination in Small Language Models (SLMs) has prompted a range of mitigation techniques. Unlike Large Language Models (LLMs), SLMs cannot rely on sheer scale to implicitly encode knowledge or reasoning paths. Instead, hallucination prevention in SLMs must be strategic, modular, and efficient, leveraging methods that compensate for their limited capacity. This section explores five key approaches used to address hallucination in SLMs.

4.1 Knowledge Grounding and Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a prominent method that integrates SLMs with external knowledge sources to enhance factual consistency during generation. In this architecture, the model retrieves relevant documents or passages from a knowledge base in real time and tailors its generation on this external context [22]. This decouples the need to store factual knowledge within model parameters—an especially critical adaptation for SLMs with constrained capacity.

For SLMs, retrieval modules are typically lightweight and may rely on sparse retrievers (e.g., BM25) or dense retrievers (e.g., DPR or ColBERT) [45–47]. When combined with efficient attention mechanisms (e.g., late fusion or shallow concatenation), RAG allows SLMs to produce more factually grounded outputs without expanding model size [22].

One implementation strategy involves embedding the retriever inside the inference pipeline, so each query dynamically retrieves the most relevant contexts. Studies have shown that even SLMs under 100M parameters can benefit from plug-and-play grounding, reducing hallucination rates in QA and summarization by 20–30% [19, 48].

However, RAG introduces dependencies on retrieval accuracy and latency, which may limit real-time applications or those operating in offline environments.

4.2 Instruction Tuning and Prompt Engineering

Instruction tuning is a method where models are trained to follow task-specific prompts formulated as natural language instructions [49]. For SLMs, this technique enhances alignment with user intent and improves model behavior across tasks where factual accuracy is paramount.

By carefully designing prompts—either manually or through automated prompt optimization—practitioners can steer the SLM toward safer, less hallucination-prone outputs. For example, explicitly instructing the model to “only answer if certain” or “cite supporting facts” can trigger internal constraints that limit speculative generation.

Instruction tuning can be implemented through multi-task fine-tuning on datasets like FLAN, which include thousands of labeled instruction-following examples [50, 51]. For SLMs, this tuning process often yields disproportionate gains, effectively compressing task knowledge into fewer parameters and reducing hallucination without architectural changes.

Prompt engineering, on the other hand, is a zero-shot or few-shot technique where hallucination is mitigated at inference time by controlling the prompt format, context framing, and output constraints [52]. Techniques such as prompt chaining, structured templates, and declarative framing have shown to reduce hallucination frequency, especially in summarization and QA tasks.

Nevertheless, instruction tuning requires access to curated datasets, and prompt engineering may lack robustness across diverse domains.

4.3 Fact-Checking and Post-Generation Verification

Another viable strategy for hallucination prevention is to evaluate and filter model outputs through automated fact-checking systems. These systems can operate as post-processing layers that assess whether a generated statement is supported by available evidence [53].

For SLMs, lightweight classifiers or entailment models—fine-tuned on verification datasets like FEVER or SciFact—can be used to score the factual consistency of generated text [54, 55]. This two-pass architecture ensures that hallucinated outputs are flagged, re-ranked, or rejected before delivery to users.

Some frameworks also incorporate self-checking mechanisms, where the SLM itself is queried post-generation to verify its prior output. While promising, such methods require careful calibration to avoid recursive error propagation.

In deployment settings where SLMs serve real users, incorporating feedback-based rejection sampling—where only outputs passing factuality filters are accepted—can significantly reduce the incidence of hallucination. However, this may increase latency or reduce generation fluency if over-applied.

4.4 Calibration Techniques (e.g., Temperature Scaling, Confidence Regularization)

Calibration techniques adjust the model's output distribution to better reflect its confidence in generation, thereby reducing speculative or hallucinated completions. One common technique is temperature scaling, which adjusts the softmax temperature during decoding to control the diversity and randomness of generated outputs [56].

For SLMs, using lower temperatures typically results in more conservative and factually stable outputs, though sometimes at the cost of creativity or informativeness. Top-k or nucleus sampling (top-p) can further constrain the token space to probable completions, reducing hallucination risk.

Another promising technique is confidence regularization during training, where the model is penalized for generating high-confidence outputs that are later found to be incorrect. These approaches aim to align confidence with correctness, allowing downstream components to use confidence scores as proxies for factual reliability [27].

For instance, calibrating confidence thresholds to trigger rejection or fallback queries can empower SLMs with error-aware control logic without altering the base architecture.

4.5 Fine-Tuning with Human Feedback

Fine-tuning with human feedback represents a powerful technique for aligning SLM behavior with human expectations, especially for controlling hallucinations. The most widely studied method in this category is Reinforcement Learning from Human Feedback (RLHF), which trains the model not only to produce coherent responses, but also to prioritize outputs rated as helpful, harmless, and truthful by human annotators [57].

While RLHF has been predominantly applied to LLMs, its principles are transferable to SLMs with appropriate efficiency adaptations. Instead of full-scale reinforcement learning pipelines, SLMs can leverage simplified forms such as Direct Preference Optimization (DPO), which skips the reinforcement learning step and fine-tunes the model directly on preference pairs [58].

For hallucination prevention, the feedback signal focuses on factuality, coherence, and source fidelity. For example, annotators may rank responses based on factual correctness, enabling the model to learn implicit representations of truthfulness. Once trained, the model internalizes these preferences and exhibits significantly lower hallucination rates, even under ambiguous or underspecified prompts.

In low-resource settings where human annotation is limited, proxy signals such as automatic entailment scores or factual consistency metrics can be used to simulate feedback. This hybrid approach, though less precise than full human oversight, enables scalable alignment tuning of SLMs with factuality as a core training objective.

However, RLHF and DPO are data-intensive and sensitive to annotation quality. Improper or biased preferences can lead to overfitting or suppression of valid alternative expressions. As such, they are most effective when paired with robust evaluation protocols and diverse human raters.

5. Performance Enhancement Techniques

Due to their constrained capacity, Small Language Models (SLMs) require tailored techniques to maximize their performance across diverse tasks. Unlike Large Language Models (LLMs), which benefit from sheer scale, SLMs must leverage parameter-efficient, resource-aware strategies to reach comparable utility. This section reviews five such methods, focusing on architecture optimization, training efficiency, and knowledge transfer.

5.1 Quantization and Pruning

Quantization refers to the process of converting a model's weights and activations from high-precision (e.g., 32-bit floating point) to lower precision formats such as 8-bit, 4-bit, or even binary representations [15]. This reduces the model size and memory footprint, allowing for faster inference and deployment on edge devices without necessarily retraining the model from scratch.

SLMs are particularly suited to quantization because their smaller architectures make quantization-aware training and fine-tuning more tractable. Empirical studies have shown that well-quantized SLMs can preserve over 95% of their original performance, even in dense generative tasks [59].

Pruning, on the other hand, involves removing redundant weights, attention heads, or entire layers based on importance scores. Methods such as magnitude pruning, structured pruning, and movement pruning can reduce the computational load and inference latency. For SLMs, pruning must be applied judiciously to avoid over-pruning, which can lead to severe degradation due to their limited redundancy [60].

Together, quantization and pruning serve as key enablers for real-time, low-power applications, including chatbots, mobile summarizers, and translation systems.

5.2 Low-Rank Adaptation and Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA) is a fine-tuning technique that freezes the original model weights and injects small trainable matrices into specific layers [38]. This significantly reduces the number of trainable parameters—often by over 90%—while maintaining competitive performance across a wide range of tasks.

LoRA is particularly impactful for SLMs because it allows for task-specific fine-tuning with minimal memory and compute cost, enabling a single base model to serve multiple use cases. For instance, fine-tuning a 100M parameter model with LoRA may only require adjusting a few million parameters, making it feasible even in low-resource or on-device settings.

Other parameter-efficient tuning (PET) methods include adapters, prefix-tuning, and (IA)[3], each of which introduces small-scale modifications to model internals while preserving generalization [61–63]. These approaches not only improve modularity and task specialization, but also reduce the risk of catastrophic forgetting, making them ideal for multi-domain deployments [64]. PET techniques have been shown to outperform full fine-tuning in data-limited scenarios and exhibit strong transferability across domains, making them foundational to modern SLM pipelines [15, 19].

5.3 Knowledge Distillation from LLMs

Knowledge distillation involves training a compact student model (e.g., an SLM) to imitate the behaviors, logits, or hidden representations of a larger, pretrained teacher model (e.g., an LLM). The goal is to transfer knowledge from high-capacity models into smaller, deployable ones [65].

In SLMs, distillation can take several forms:

- Logit-based distillation, where the student learns to match the soft target distributions of the teacher.
- Intermediate representation distillation, aligning hidden states or attention maps.
- Response-based distillation, where the student mimics the outputs of the teacher in generative or classification tasks.

Modern techniques such as TinyBERT and DistilBERT have shown that carefully distilled SLMs can achieve up to 97% of the teacher's accuracy on GLUE benchmarks while being 40–60% faster [60, 66].

Distillation is especially useful when pretraining from scratch is infeasible, and when deploying models in latency-sensitive environments. However, distillation quality is highly dependent on teacher diversity, task alignment, and training signal richness.

5.4 Mixture of Experts (MoE) for SLMs

Although traditionally associated with LLMs, Mixture of Experts (MoE) architectures have recently been adapted for use in SLMs. MoE models consist of multiple expert subnetworks, with a gating mechanism that activates only a subset during inference, thereby increasing capacity without linearly increasing compute [67].

In the context of SLMs, sparse MoE architectures can be implemented by using small-scale expert modules specialized for tasks, domains, or languages. When only a few experts are active at each step, the effective model size grows, but the runtime cost remains constant. This enables SLMs to achieve LLM-level performance on specific subtasks while remaining efficient [68].

Some lightweight MoE models—such as Switch Transformers Lite—have demonstrated that even models with under 200M parameters can benefit from expert sparsity, especially in multitask learning and dialogue generation [68]. Key challenges remain in balancing expert usage and avoiding overfitting to particular experts, but the approach is promising for scalable SLM enhancement.

5.5 Specialized Training Regimes (Task-specific and Domain-adaptive)

Another impactful strategy for improving SLM performance is the use of specialized training regimes, tailored to either a specific task or a target domain. Since SLMs lack the overparameterization of LLMs, their ability to generalize broadly is constrained—making focused and deliberate training essential to maximize utility.

Task-specific fine-tuning

SLMs can be fine-tuned on carefully curated task datasets (e.g., sentiment analysis, summarization, NER) using supervised learning. When the dataset is well-matched and of sufficient quality, even a low size model can approach or exceed LLM performance on narrow tasks [69]. Fine-tuning with task-specific loss functions, such as factuality-aware or span-based objectives, further optimizes model behavior toward the desired output format.

Domain-adaptive pretraining (DAPT)

In cases where the target domain differs significantly from general corpora (e.g., legal, medical, scientific text), Domain-Adaptive Pretraining helps by continuing pretraining on unlabeled in-domain data before fine-tuning. This enables the SLM to learn domain-specific vocabulary, syntax, and concepts that would otherwise be underrepresented in general training data [70].

DAPT is particularly effective for SLMs because it allows for efficient specialization without complete retraining. In biomedical applications, for example, BioBERT and SciBERT-style adaptations have produced significant gains using only modestly sized models [40, 71].

Multi-task and Curriculum Learning

Advanced regimes also include multi-task learning, where the model is trained simultaneously on related tasks (e.g., QA + summarization), and curriculum learning, where training progresses from simple to complex examples. These techniques help mitigate overfitting and promote better generalization—critical properties for underparameterized SLMs [72].

In all cases, the success of these regimes depends on the alignment between training data and deployment scenarios. SLMs are more brittle than LLMs under distribution shift, and thus require more careful data selection and evaluation.

6. Comparative Analysis

This section synthesizes the techniques introduced in previous discussions by conducting a comparative analysis in terms of their effectiveness, computational efficiency, scalability, and applicability to real-world deployment of Small Language Models. Given the diversity of methods available, understanding which to use, when, and how to combine them is essential for researchers and engineers seeking to maximize both factual accuracy and functional performance under limited resources. Table 1 compares considered hallucination prevention and performance enhancement methods

6.1 Resource Requirements vs. Performance Gains

The utility of a method often hinges on the performance gain it offers relative to its cost in resources—a central concern in SLM development. Techniques like LoRA, quantization, and instruction tuning strike a particularly favorable balance. LoRA enables up to 90% reduction in trainable parameters with minimal performance loss [38], while 8-bit quantization can reduce memory footprint by 4× with <2% degradation in accuracy [59].

By contrast, methods such as RLHF and RAG yield high returns in hallucination reduction but incur substantial compute and human annotation overheads. RLHF, for instance, requires thousands of preference-labeled instances and multiple optimization stages [57]. RAG's dependence on real-time retrieval and memory access may disqualify it from edge-device deployments.

Table 1. Comparison of considered hallucination prevention and performance enhancement methods

Method	Category	Advantages	Disadvantages	Typical Use Case
Retrieval-Augmented Generation (RAG)	Hallucination Prevention	Strong grounding; reduces factual errors; domain adaptable	Requires external retriever and memory; slower inference	QA, document summarization, open-domain tasks
Instruction Tuning	Hallucination Prevention	Improves alignment and intent-following; reusable across tasks	Needs curated datasets; sensitive to prompt phrasing	Multi-task systems, APIs, general assistants
Fact-checking/ Verification	Hallucination Prevention	Adds post-hoc quality control; modular	Increases latency; may miss subtle hallucinations	Critical domains (e.g., medical, legal)
Calibration Techniques	Hallucination Prevention	Controls randomness; enhances output reliability	Reduces diversity; requires tuning	Controlled generation; high-stakes applications
RLHF / DPO	Hallucination Prevention	Aligns outputs to human preferences; adaptable across models	High annotation cost; risk of reward hacking	Safety-critical dialogue, personal assistants
Quantization & Pruning	Performance Enhancement	Reduces memory and compute; enables edge deployment	Can degrade accuracy if aggressive	Real-time applications, mobile devices
LoRA / Parameter-Efficient Tuning	Performance Enhancement	Minimal resource use; enables fast adaptation	Adds complexity to training pipeline	Domain-specific or multi-task settings
Knowledge Distillation	Performance Enhancement	Transfers LLM knowledge into compact models; scalable	Dependent on teacher quality; less adaptable post-distillation	Model compression, enterprise deployments
Mixture of Experts (MoE)	Performance Enhancement	Expands capacity without linear compute growth	Complex routing; hard to train and balance	Specialized agents, multitask systems
Domain-Adaptive & Task-Specific Tuning	Performance Enhancement	Highly effective in matched domains; easy to implement	Narrow generalization; requires curated data	Biomedical NLP, legal analysis, customer service

In low-resource or latency-sensitive environments, quantization, fact-checking with lightweight classifiers, and prompt engineering are often the only viable options. Meanwhile, infrastructure-rich settings (e.g., cloud services, research labs) can afford to implement more intensive techniques like RLHF or dynamic MoE routing for SLM orchestration.

6.2 When to Use What: Use-Case-Driven Combinations

Selecting and combining hallucination prevention and performance enhancement techniques depends heavily on the deployment context, available resources, and task criticality. While individual methods offer isolated benefits, their synergistic use often yields optimal performance. The following scenarios illustrate combinations of methods best suited for typical SLM applications:

- Mobile or On-Device NLP Applications: For environments with severe compute and memory constraints, a combination of 8-bit quantization, structured pruning, and parameter-efficient fine-tuning (e.g., LoRA) allows for real-time inference with minimal loss in performance [15, 38]. Temperature tuning can be applied to suppress hallucinated content while preserving fluency.
- Enterprise Question Answering Systems: Applications such as legal or technical support benefit from retrieval-augmented generation (RAG) combined with domain-adaptive pretraining [70] and instruction tuning [73]. For high-fidelity tasks, integrating post-hoc verification improves factual reliability, as shown in medical summarization and scientific QA contexts [53].
- Multi-Task Digital Assistants: Assistant models designed for varied conversational goals should employ LoRA or adapters for modular task specialization [64], augmented by multi-task instruction tuning and direct preference optimization (DPO) to reduce hallucination and align outputs with user expectations [58].
- Safety-Critical Applications (e.g., Healthcare, Finance): In these domains, hallucination poses legal and ethical risks. Here, low-temperature decoding, confidence regularization, and factuality-based filtering pipelines are essential. Fine-tuning with human-labeled preference datasets [57] further ensures output safety.
- Content Creation and Education Tools: These applications may prioritize creativity and fluency but still require factual grounding. Distillation from high-performing LLMs and structured prompt engineering are effective here [14, 52, 66]. When available, RAG modules improve informativeness with minimal hallucination.

Each of these combinations showcases the adaptability of SLMs when engineering constraints and task-specific needs are jointly considered.

6.3 Scalability and Portability Across Platforms and Domains

The scalability and portability of hallucination prevention and performance enhancement methods vary significantly depending on their algorithmic structure and infrastructure dependency.

- Highly Portable Methods: Techniques such as quantization, LoRA, and instruction tuning are architecturally non-invasive and computationally efficient, making them suitable for cross-platform deployment, including mobile, edge, and embedded devices [15, 38, 49]. Their lightweight implementation also facilitates deployment in environments with intermittent connectivity or offline constraints.
- Limited Portability Techniques: Strategies like retrieval-augmented generation, reinforcement learning from human feedback (RLHF), and mixture-of-experts

(MoE) are resource-intensive and require supporting infrastructure (e.g., real-time retrieval databases, expert gating mechanisms, or human annotation loops). While effective in high-resource settings, they are challenging to scale to decentralized or embedded systems [57, 68].

- Cross-Domain Generalizability: Methods like instruction tuning, distillation, and multi-task fine-tuning have shown high generalizability across NLP benchmarks and languages [66, 74]. However, domain-adaptive pretraining remains critical for performance in specialized fields such as medicine or law, where general-domain pretraining is insufficient [70].

In summary, method selection should account for not only model size and accuracy, but also operational constraints such as target hardware, latency tolerance, and domain specificity. A modular, composable approach offers the best path toward scalable, efficient, and reliable SLM deployment.

7. Discussion

One of the most pressing challenges in optimizing Small Language Models (SLMs) is achieving compatibility among multiple performance and hallucination mitigation strategies. As shown in the prior sections, many techniques target different dimensions of model behavior—some modify the base architecture (e.g., pruning, LoRA), while others affect training procedures (e.g., instruction tuning, RLHF), and still others operate externally (e.g., fact-checkers, RAG modules).

Successfully integrating these approaches requires attention to their interdependencies and interference potential. For example, deploying quantization alongside RAG may lead to mismatches between token embeddings and retrieval vectors unless carefully aligned [15]. Similarly, combining LoRA with RLHF necessitates tuning not only reward signals but also how and where LoRA adapters are applied—especially in transformer attention layers [38, 58].

There is also the question of interoperability across tools and pipelines. For instance, many fine-tuning and alignment strategies assume access to open-weight models and fine-grained control over layers. This poses a challenge when using proprietary or API-constrained SLMs, where only prompt-level tuning is possible. In such cases, prompt engineering and calibration become the only feasible levers, despite their lower ceiling on performance.

Thus, the future of effective SLM deployment likely lies in modular, plug-and-play systems that allow for composable combinations of retrieval, tuning, and filtering techniques, each calibrated to the resource and risk profile of the application.

7.1 Sustainability and Cost-Efficiency Considerations

While much of the focus in NLP research has been on maximizing performance, the rising emphasis on sustainability and cost-efficiency compels a re-evaluation of model design priorities. Large Language Models have been criticized for their carbon footprint, hardware demands, and data inefficiency [46]. In contrast, SLMs offer a promising alternative—but only if their enhancement techniques are themselves sustainable.

Many of the most effective hallucination prevention and performance boosting methods—such as RLHF, RAG, or MoE—require significant training or infrastructure overhead, which can offset the lightweight benefits of the base model. A key direction for future work involves developing low-cost analogues of these methods that retain most of the performance gains. For example, pseudo-labeling for alignment or factuality-aware pretraining might provide lightweight alternatives to human-in-the-loop approaches [21].

Moreover, SLM-centric design needs to shift toward lifecycle-aware modeling, where training, deployment, and continual learning are planned with compute budgets in mind. Techniques such as progressive freezing, dynamic fine-tuning, and on-device continual learning could allow for long-term use with minimal retraining [75].

7.2 Ethical Considerations and Transparency

The deployment of SLMs—especially in high-stakes and consumer-facing contexts—raises important ethical questions, particularly around trust, transparency, and safety. Even though SLMs are less capable than LLMs, they can still produce hallucinated content that users mistake for truth, particularly due to their fluent style and confident tone [13]. This risk is magnified in applications involving children, vulnerable populations, or marginalized groups.

One promising mitigation strategy is to embed factual confidence scores or disclaimers into model outputs—an approach shown to improve user calibration and reduce overreliance on model suggestions [27]. Moreover, transparency about model capabilities and limitations, including accuracy rates and hallucination tendencies, should be a standard component of model documentation [76].

From a governance perspective, organizations deploying SLMs should consider auditable logs of model decisions, especially when using post-generation filtering or human feedback integration. This is critical for ensuring accountability in regulated domains such as healthcare, finance, and education.

Finally, the development of benchmark suites tailored for SLMs, particularly around factuality and task alignment, is needed. Current evaluation standards disproportionately favor LLMs and do not adequately reflect the real-world constraints or behaviors of smaller models.

7.3 Future Directions in SLM Optimization

Several open research directions emerge from this review:

- SLM-centric pretraining paradigms: Rather than distilling from LLMs or truncating larger architectures, future models could be trained from scratch using SLM-optimized curricula that prioritize factual grounding, compression efficiency, and domain relevance.
- Neural-symbolic hybrids: Combining compact language models with symbolic reasoning engines or structured databases may offer improved reliability, especially in logic-heavy or tabular domains [77].
- Benchmarking under constraints: New evaluation datasets and leaderboards should be created that explicitly account for compute, memory, and latency limits, promoting fair and transparent comparisons of SLM capabilities.
- Factuality-aware decoding strategies: Advances in controllable generation could empower SLMs to internally regulate hallucinations during output generation, obviating the need for post-processing in many contexts.

Overall, the path forward involves rethinking NLP not only as a scale game, but as a systems optimization problem—balancing factuality, efficiency, and trust in constrained environments.

8. Conclusion

This review has examined the hallucination prevention and performance enhancement techniques applicable to Small Language Models (SLMs), with an emphasis on their practical deployment in resource-constrained and real-world settings. In contrast to the dominant narrative centered on Large Language Models (LLMs), this paper highlights that SLMs—despite their limitations in capacity—can be made competent, trustworthy, and efficient through the careful application of modular and synergistic strategies.

For hallucination mitigation, five major approaches were surveyed: retrieval-augmented generation (RAG), instruction tuning and prompt design, post-hoc fact-checking, calibration methods, and preference-based alignment such as RLHF or DPO. Each offers unique strengths and trade-offs. For performance enhancement, methods such as quantization, pruning, parameter-efficient tuning (e.g., LoRA), knowledge distillation, MoE architectures, and domain-adaptive training were reviewed. These strategies have demonstrated substantial gains in both efficiency and task performance, even in low-resource or real-time inference environments.

In light of the comparative analysis no single method seems universally optimal. Rather, method selection should be driven by deployment requirements, such as latency, memory constraints, and domain specificity. Effective SLM pipelines often involve hybrid configurations, balancing accuracy with computational feasibility.

Based on the findings, we offer the following recommendations for practitioners and researchers working with SLMs:

- For On-Device Deployment: Prioritize quantization, pruning, and parameter-efficient fine-tuning to balance speed and accuracy. Use calibration techniques like temperature scaling to manage hallucination risk without incurring additional compute.
- For Domain-Specific Applications: Apply domain-adaptive pretraining and instruction tuning. Where feasible, integrate lightweight retrieval modules or task-specific factuality filters.
- For Critical Systems (e.g., medical, legal): Combine instruction tuning with output verification modules. Prefer conservative decoding strategies and, when possible, introduce RLHF or DPO to enforce user-aligned safety.
- For Scalable Multi-Task Systems: Use LoRA with shared base models and task-specific adapters. Consider prompt standardization or instruction chaining to reduce hallucination and improve generalization.
- For Model Development and Research: Invest in benchmarks and metrics tailored to SLMs—especially ones that emphasize factuality, robustness, and cost-performance balance.

One of the important gaps in the current research landscape is the lack of standardized, publicly available benchmarks for evaluating SLMs on both factuality and efficiency metrics. Existing evaluation protocols tend to overfit the needs of LLMs, emphasizing scale-driven generalization and zero-shot capabilities. These metrics fail to capture the pragmatic trade-offs that dominate SLM usage—such as inference speed, memory use, and factual error rates in low-capacity regimes.

SLM-specific evaluation suites that incorporate the following may be beneficial for further advancement of the field:

- Task coverage diversity, including both generative and discriminative benchmarks;

- Multi-domain factuality tests, ideally linked to human-annotated ground truths;
- Efficiency-performance trade-off metrics, measuring energy consumption, latency, and model size against output quality;
- Robustness to prompt variation and domain drift, assessing stability and hallucination under real-world noise.

The development of such benchmarks may be regarded as important to advancing the responsible and effective deployment of SLMs in academic, commercial, and public-interest settings.

References

[1] Liddy ED. Natural language processing. In: Encyclopedia of Library and Information Science. 2nd Ed. New York: Marcel Decker, Inc.; 2001.

[2] Jurafsky D, Martin JH. *Speech and Language Processing*. Pearson; 2023.

[3] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 http://arxiv.org/abs/1301.3781.

[4] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–43.

[5] Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 2227–37.

[6] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 4171–86.

[7] Radford A, Narasimhan K, Salimans T, Sutskever I, others. Improving language understanding by generative pre-training. 2018;

[8] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019; 1(8):9.

[9] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. p. 1877–901.

[10] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways. 2022 http://arxiv.org/abs/2204.02311.

[11] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report. 2023 http://arxiv.org/abs/2303.08774.

[12] Alizadeh K, Mirzadeh SI, Belenko D, Khatamifard S, Cho M, Del Mundo CC, et al. LLM in a flash: Efficient large language model inference with limited memory. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024. p. 12562–84.

[13] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023; 55(12):1–38.

[14] Turc I, Chang M-W, Lee K, Toutanova K. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. 2019 http://arxiv.org/abs/1908.08962.

[15] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. 2023 http://arxiv.org/abs/2305.14314.

[16] Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational

Linguistics; 2020. p. 1906–19.

[17] Naveen B, Ashita S, Pushpak B. NLI to the Rescue: Mapping Entailment Classes to Hallucination Categories in Abstractive Summarization. In: Proceedings of the 20th International Conference on Natural Language Processing (ICON). 2023. p. 120–32.

[18] Honovich O, Aharoni R, Herzig J, Taitelbaum H, Kukliansy D, Cohen V, et al. TRUE: Re-evaluating Factual Consistency Evaluation. In: Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering. Stroudsburg, PA, USA: Association for Computational Linguistics; 2022. p. 161–75.

[19] Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval Augmentation Reduces Hallucination in Conversation. 2021 http://arxiv.org/abs/2104.07567.

[20] Chae K, Choi J, Jo Y, Kim T. Mitigating Hallucination in Abstractive Summarization with Domain-Conditional Mutual Information. In: Findings of the Association for Computational Linguistics: NAACL. 2024. p. 1809–20.

[21] Huang S, Dong L, Wang W, Hao Y, Singhal S, Ma S, et al. Language is not all you need: Aligning perception with language models. Adv Neural Inf Process Syst. 2023; 36:72096–109.

[22] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020; 33:9459–74.

[23] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020; 21(140):1–67.

[24] Olshaker H, Brin D, Kalderon E, Kraus M, Konen E, Klang E. Evaluating the Diagnostic Performance of Large Language Models in Identifying Complex Multisystemic Syndromes: A Comparative Study with Radiology Residents. 2024.

[25] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digit Med. 2021; 4(1):31.

[26] Ghweeba M, Lindenmeyer A, Shishi S, Waheed A, Kofi M, Amer S. The Attitudes of Egyptian Web-Based Health Information Seekers Toward Health Information Provided Through the Internet: Qualitative Study. JMIR Form Res. 2022; 6(2):e30108.

[27] Kadavath S, Conerly T, Askell A, Henighan T, Drain D, Perez E, et al. Language Models (Mostly) Know What They Know. 2022 http://arxiv.org/abs/2207.05221.

[28] Fabbri AR, Kryściński W, McCann B, Xiong C, Socher R, Radev D. Summeval: Re-evaluating summarization evaluation. Trans Assoc Comput Linguist. 2021; 9:391–409.

[29] Kryściński W, McCann B, Xiong C, Socher R. Evaluating the Factual Consistency of Abstractive Text Summarization. 2019 http://arxiv.org/abs/1910.12840.

[30] Laban P, Schnabel T, Bennett PN, Hearst MA. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Trans Assoc Comput Linguist. 2022; 10:163–77.

[31] Lin S, Hilton J, Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. 2021 http://arxiv.org/abs/2109.07958.

[32] Weidinger L, Uesato J, Rauh M, Griffin C, Huang P-S, Mellor J, et al. Taxonomy of risks posed by language models. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 2022. p. 214–29.

[33] Kim G, Cho K. Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search. 2020 http://arxiv.org/abs/2010.07003.

[34] Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, et al. Aligning Large Language Models with Human: A Survey. 2023 http://arxiv.org/abs/2307.12966.

[35] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022; 35:27730–44.

[36] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. 2019

http://arxiv.org/abs/1909.11942.
[37] Schick T, Schütze H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 2339–52.
[38] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. 2021 http://arxiv.org/abs/2106.09685.
[39] Zhang W, Hou L, Yin Y, Shang L, Chen X, Jiang X, et al. TernaryBERT: Distillation-aware Ultra-low Bit BERT. 2020 http://arxiv.org/abs/2009.12812.
[40] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, editor. Bioinformatics. 2020; 36(4):1234–40.
[41] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. 2019 https://arxiv.org/abs/1904.03323.
[42] Grassi L, Recchiuto CT, Sgorbissa A. Knowledge-grounded dialogue flow management for social robots and conversational agents. Int J Soc Robot. 2022; 14(5):1273–93.
[43] Van Nguyen C, Shen X, Aponte R, Xia Y, Basu S, Hu Z, et al. A Survey of Small Language Models. 2024 http://arxiv.org/abs/2410.20011.
[44] Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: Self-Attention with Linear Complexity. 2020 http://arxiv.org/abs/2006.04768.
[45] Robertson S, Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. Found Trends® Inf Retr. 2009; 3(4):333–89.
[46] Khattab O, Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2020. p. 39–48.
[47] Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 6769–81.
[48] Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. 2020 http://arxiv.org/abs/2007.01282.
[49] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned Language Models Are Zero-Shot Learners. 2021 http://arxiv.org/abs/2109.01652.
[50] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling Instruction-Finetuned Language Models. 2022 http://arxiv.org/abs/2210.11416.
[51] Longpre S, Hou L, Vu T, Webson A, Chung HW, Tay Y, et al. The flan collection: Designing data and methods for effective instruction tuning. In: International Conference on Machine Learning. 2023. p. 22631–48.
[52] Wang Y, Wang M, Manzoor MA, Liu F, Georgiev G, Das RJ, et al. Factuality of Large Language Models: A Survey. 2024 http://arxiv.org/abs/2402.02420.
[53] Atanasova P. Generating fact checking explanations. In: Accountable and Explainable Methods for Complex Reasoning over Text. Springer; 2024. p. 83–103.
[54] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 809–19.
[55] Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or Fiction: Verifying Scientific Claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 7534–50.

[56] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International conference on machine learning. 2017. p. 1321–30.

[57] Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 2017; 30.

[58] Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. 2023 http://arxiv.org/abs/2305.18290.

[59] Frantar E, Ashkboos S, Hoefler T, Alistarh D. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. 2022 http://arxiv.org/abs/2210.17323.

[60] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2020 https://arxiv.org/abs/1910.01108.

[61] Li XL, Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 4582–97.

[62] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. Parameter-efficient transfer learning for NLP. In: International conference on machine learning. 2019. p. 2790–9.

[63] Liu H, Tam D, Muqeeth M, Mohta J, Huang T, Bansal M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Adv Neural Inf Process Syst. 2022; 35:1950–65.

[64] Mahabadi RK, Ruder S, Dehghani M, Henderson J. Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks. 2021 http://arxiv.org/abs/2106.04489.

[65] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. 2015 http://arxiv.org/abs/1503.02531.

[66] Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, et al. TinyBERT: Distilling BERT for Natural Language Understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 4163–74.

[67] Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. 2017 http://arxiv.org/abs/1701.06538.

[68] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res. 2022; 23(120):1–39.

[69] Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 328–39.

[70] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. 2020 http://arxiv.org/abs/2004.10964.

[71] Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. 2019 http://arxiv.org/abs/1903.10676.

[72] Xu B, Zhang L, Mao Z, Wang Q, Xie H, Zhang Y. Curriculum learning for natural language understanding. In: Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. p. 6095–104.

[73] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. 2022 http://arxiv.org/abs/2205.11916.

[74] Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafeai Z, et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. 2021 http://arxiv.org/abs/2110.08207.

[75] Aghajanyan A, Shrivastava A, Gupta A, Goyal N, Zettlemoyer L, Gupta S. Better Fine-Tuning by Reducing Representational Collapse. 2020 http://arxiv.org/abs/2008.03156.

[76] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the conference on fairness, accountability, and transparency. 2019. p. 220–9.

[77] Chen X, Liang C, Yu AW, Zhou D, Song D, Le Q V. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In: International Conference on Learning Representations. 2019.

Research & Design

In This Issue

Research Group

MIM Research & Publishing