Filtering hallucinations and omissions in Large Language Models through a cognitive architecture

Md Asaduzzaman
School of Computing
University of Kent
Canterbury, United Kingdom
ma2136@kent.ac.uk

Ioanna Giorgi
School of Computing
University of Kent
Canterbury, United Kingdom
i.giorgi@kent.ac.uk

Giovanni L Masala*
School of Computing
University of Kent
Canterbury, United Kingdom
g.masala@kent.ac.uk

***Abstract*—While Large Language Models (LLMs) have outpaced recent technological advancements, challenges like hallucinations and omissions persist in all LLMs due to the underlying architecture and model training. Hallucinations refer to instances where the model generates incorrect, fabricated, or ungrounded information. Omissions occur when the model fails to provide certain details or skips relevant information in its response.**

This paper proposes a novel hybrid methodology to mitigate these phenomena by integrating an LLM (GPT-3.5) with an external brain-inspired cognitive architecture. Unlike classical approaches, our hybrid system leverages mechanisms for long-term memory, structured reasoning, and multi-modal learning and presents further opportunities for improving LLMs with continuous learning, multilingual skills and focus of attention, without ad hoc fine-tuning. The hybrid system was tested and compared with two standalone LLMs (GPT-3.5, Gemini) through simulated open dialogue that mimic daily conversations These tests involved implicit conversational questions or statements on topics like social contexts and basic knowledge, e.g., discussing animals or comparing numbers.

Our proposed model reduced hallucinations and omissions compared to the standalone LLMs on the same benchmark dataset. Specifically, reductions were observed in (i) hallucinations: 33.85% over GPT-3.5 and 37.48% over Gemini, (ii) omissions: 29.80% over GPT-3.5 and 27.20% over Gemini; (iii) instruction loss: 8.13% over GPT-3.5 & 4.68% over Gemini.

Keywords—LLM, hallucinations, omissions, cognitive models

I. Introduction

Large Language Models (LLMs), such as GPT-3 [1] and its successors, have transformed natural language processing (NLP) by generating human-like text, engaging in complex dialogue, and performing a wide range of linguistic tasks. However, they face significant challenges, particularly hallucinations and omissions. Hallucinations occur when the model generates factually incorrect or fabricated information, while omissions involve the exclusion of relevant details, leading to incomplete or misleading responses [2]. These issues are critical concerns for the reliability and usability of LLMs, particularly in fields like healthcare, legal advisory, and education, where accuracy and completeness are essential. The authors in [3] highlight the significance of LLM hallucinations and omission in the meteorology domain classifying them by severity with the help of an expert in such domain. Currently, there is no established clinical framework for assessing the safety and accuracy of LLM-generated medical text [4]. For example, authors in [5] utilised LLMs for creating clinical documentation, showing the potential of these models in generating medical reports. However, assessing the reliability of these generated texts was beyond the focus of their study. Authors in [4] categorise LLM errors within the clinical documentation context, establishing some metrics and a framework for assessing the safety risk of errors. Fine-tuning LLMs specifically for these domains may not necessarily enhance model performance and could lead to a spike in hallucinations when generating new content [6].

There exist attempts to mitigate hallucinations in LLMs. Article [7] aims to minimise hallucinations in LLM text generation by incorporating contextual data about relevant entities mapped through a knowledge graph. In the comprehensive survey of hallucination mitigation techniques in LLMs [8], authors identify over 32 techniques developed to mitigate this issue. Some notable approaches include Retrieval Augmented Generation [9], Knowledge Retrieval [10], CoNLI [11], and CoVe [12]. Typically the main effort is to improve the detection of the hallucination using mitigation techniques that do not introduce new hallucinations.

In this paper, we propose a new paradigm to address hallucinations and omissions in LLMs by designing a *hybrid system* that leverages a brain-inspired cognitive architecture alongside the LLM. Unlike traditional techniques that focus on detecting and correcting errors in real-time, a cognitive architecture possesses mechanisms for humanlike processes like long-term memory, structured reasoning and inference, which enhance context understanding, information retention, and inference accuracy. Our approach provides opportunities to develop a high-cognitive and versatile conversational agent capable of continuous learning and multilingual cognition.

II. Rationale for Cognitive Architecture

Hallucinations in LLMs primarily result from their design as inference engines that produce "best guess" responses from statistical likelihoods. To reduce such stochastic parroting, our methodology is informed by language production processes of the human brain. These are computationally modelled in cognitive architectures that hypothesise core brain structures, modules, and their relationships [13]. Cognitive architectures, (e.g., ACT-R [14], SOAR [15]), are used to derive intelligent behaviour in complex environments by reverse-engineering aspects of human cognition, like memory, perception, and reasoning, for artificial general intelligence (AGI) [16].

A significant subset of cognitive architectures focuses on psychological principles [17]. The underpinning mechanisms and processes are usually defined in sufficient detail for computational implementation and grounded on the same cognitive primitives that can explain ample tasks, data, and phenomena, making their scope more generic. The rationale for choosing a cognitive architecture, over some domain-specific memory source for LLMs or fine-tuning them in closed domains, is thus that of being an antithesis to narrow systems. Also, cognitive architectures provide explanations based on the underlying cognitive mechanisms, which more closely approximate the theory of mind [17]. This makes them

valuable for interpreting the decision processes of the LLMs' black-box learning algorithms. While LLMs aim for optimal solutions based on data-heavy training, cognitive architectures prioritise rational decision-making amid resource constraints.

A key notion in large cognitive architectures is working memory (WM), which plays a crucial role in cognitive tasks by retaining relevant information and excluding irrelevant stimuli [18]. In contrast, LLMs rely on fixed context windows that discard older tokens once the limit is reached, potentially causing context loss. While advanced models like Gemini 1.5 Pro offer a context window of 2M tokens [19], this comes with quadratic growth in parameters and higher computational costs. WM mechanisms selectively and intentionally recall, store, and manipulate information between long-term and short-term memory, thus enabling human-like capabilities such as learning, reasoning, abstraction, and memory [18]. Cognitive architectures based on WM principles can better simulate human cognition, without fine-tuning [20]. Hence, a hybrid system, despite a more complex global organisation, may achieve better context retention with less computational overhead compared to simple expanding context windows.

This paper's methodology uses the ANNABELL cognitive architecture (Artificial Neural Network with Adaptive Behaviour Exploited for Language Learning) [20], which implements Baddeley's working memory model [21] to explain how procedural knowledge can be developed in neural networks using biologically motivated learning rules. The model has demonstrated abilities for continuous runtime learning, knowledge inference, categorisation, and abstract concept formation without retraining or fine-tuning [22], thus supporting the notion that its cognitive plausible mechanisms are applicable across various tasks and data. Other studies [23] have shown its ability to maintain coherence in multilingual dialogues, that could be leveraged in multilingual LLMs.

III. Methods

The proposed hybrid system integrates OpenAI GPT with the ANNABELL cognitive architecture. For comparison, two standalone LLMs, GPT and Google Gemini, were trained on the same data. A custom user interface allows interaction in three modes: (i) GPT only, (ii) Gemini only, (iii) the hybrid GPT+ANNABELL system. In this setup, all three LLM-based systems, except the cognitive component of the hybrid system, are prompt-engineered using chain-of-thought prompting.

A. GPT and Gemini

We employed Gemini 1.5 Flash [19] and GPT-3.5 [24], both transformer-based architectures [25]. GPT-3.5 is an autoregressive model based on a Generative Pre-Trained Transformer. It uses multiple layers, attention mechanisms, and deep learning techniques to process and generate real-time text. Trained on large datasets from web pages, books, Wikipedia, and user-generated content, it has 175 billion parameters and a context window of 16,385 tokens [24]. Gemini is a decoder-only transformer model designed for multimodal processing (text, images, and other data types). It handles tasks like text summarisation, classification, content understanding, and object recognition. Gemini 1.5 Flash has 70B parameters and a context window of 1M tokens [19]

Both models support fine-tuning, but this is highly reliant on task-specific fine-tuning using datasets with thousands, or even tens of thousands, of examples. In contrast, humans can typically learn a new language task from only a few examples or simple instructions—an ability that current NLP systems still struggle with [26]. Moreover, high-quality, non-scarce and unbiased data and sufficient data variety are required to prevent overfitting, which may be challenging and resource-intensive. For this work, prompt-tuning was preferred, as it offers greater flexibility and is less resource-demanding. We leveraged the LLMs' existing ability to generate humanlike conversations given a common conjecture that LLMs acquire most knowledge during their pre-training [27].

B. The ANNABELL architecture

ANNABELL is a large-scale neural network with over 2 M interconnected artificial neurons structured in different subnetworks. These consist of learnable input connections that are modified through the Hebbian learning rule combined with the k-winner-take-all rule [20]. The model is built up of four main components represented in Fig. 1.

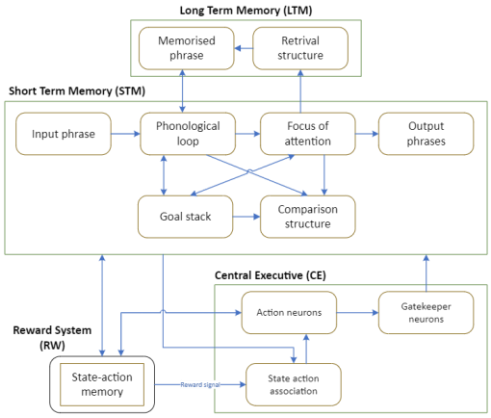

Fig. 1. Schematic diagram of the ANNABELL model (*adapted from [20]*).

The **Short-Term Memory (STM)** includes several subcomponents: a phonological store, a focus of attention, a goal stack and a comparison structure. The phonological loop handles the working phrase. The focus of attention focuses on a few words (up to four) while processing the working phrase, which also serves as a cue to retrieve information from long-term memory. The goal stack holds goal chunks for decision-making, storing actions that cannot be immediately performed but are retained as goals. The comparison structure identifies similarities between words in the focus of attention, phonological store and goal stack, to produce valid answers. The **Long-Term Memory (LTM)** stores memorised phrases and retrieval structures using cues from the focus of attention. The **Central Executive (CE)** is a supervised system that manages all decision-dependent processes through the neural gating mechanism, similar to that in the brain cortex [28]. The **Reward Structure (RW)** memorises state-action sequences from the exploration phase and rewards associations between internal states and corresponding mental actions. Thus, it links actions that produce valid answers to the activated system states during those actions. This enables generalisation and retrieval of the same state for similar inputs, using the same procedural mechanisms (mental actions). Thus, the system learns to perform actions that process data, rather than learning the data itself.

C. Vector Embedding

Three vector embedding techniques were tested for context retrieval: (i) BERT (Bidirectional Encoder Representations from Transformers) [29], (ii) a Transformer-based model of Google, OpenAI Embedding (model: text-embedding-ada-002) [30] and, (iii) TF-IDF (Term Frequency-Inverse Document Frequency) Embedding [31].

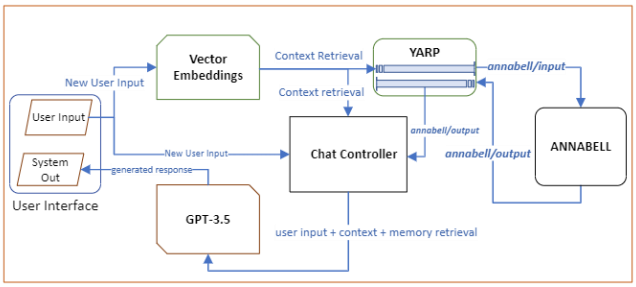

Fig. 2. The proposed hybrid system integrating GPT-3.5 with the ANNABELL cognitive architecture.

D. *The hybrid system (GPT + ANNABELL)*

The proposed hybrid system is illustrated in Fig. 2. Each user input was passed through vector embedding to convert it into vector form. This step aimed to retrieve significant words from the input context and generate relevant memory-exploitation data for the cognitive architecture. The data was fed into the architecture via the YARP (Yet Another Robot Platform) interface [32]. Querying the cognitive architecture first with contextual data from the user input was intentionally done to produce reasoning and memory-based responses that could then inform the answers generated by GPT.

The Chat Controller combines the user input, context from vector embedding, and output from the cognitive architecture into a single prompt, which is fed to GPT-3.5. This approach offers two key advantages. It isolates the significant context from the rest of the prompt tokens, focusing GPT's "attention" on the desired context during generation. It uses the cognitive model's statistical decision processes for accurate memory exploitation, discouraging GPT from hallucinating or omitting information. The final GPT-generated response from the engineered prompt is then delivered to the user.

IV. Dataset and Testing

The dataset used in this work is sourced from ANNABELL's publicly trained repository [20]. It is structured in five thematic groups, each containing both declarative (factual) and interrogative sentences (procedural knowledge). This dataset was selected as a baseline since the cognitive architecture was fully validated on it (1587 total questions), and is suitable for simulating daily conversations. The dataset also includes examples of question answers, thus supporting prompt-tuning of standalone LLM models. With a total length of 4176 tokens, the dataset exceeds the maximum content window of both LLMs.

Four user-system dialogues were used to evaluate the proposed system against standalone LLMs. The conversations focused mainly on the "People" thematic group, based on the Language Development Survey [33]. This group details the social environment of a fictional four-year-old girl, Annabell, her relationships with twenty people, and their likes, dislikes, routines, and professions. Questions from the Categorisation thematic group were also randomly included. Each test was conducted in three modes: GPT only, Gemini only and GPT+ANNABELL and the system responses were analysed for hallucinations, omissions, and loss of instruction. Our implementation also controlled the total tokens in the prompt to ensure they stayed within the maximum context window limits of the LLMs during the tests. A final test with counting and comparing numbers is also performed to compare the standalone GPT and the proposed hybrid system, as GPT-3.5 is renowned for underperforming in arithmetic calculations.

V. Results

First, we compare the results of the vector embeddings to retrieve relevant context. This is essential to ensure that the procedural knowledge of the cognitive architecture is properly harnessed. The results are illustrated with two user inputs as shown in Fig. 3. Each model is tuned to pick the top five scored sentences produced by the embedding. The diamond-grid bars represent close-context retrieval and diagonal-grid bars illustrate incorrect retrieval. OpenAI Embedding and TF-IDF yielded similar results, while BERT was suboptimal. Nevertheless, OpenAI Embedding requires storing vectors to reduce costly and time-consuming API calls. Thus, we selected TF-IDF as it delivers comparable results but is faster, supports on-premises implementations, and incurs no cost or time delay, making it more suitable for real-time applications.

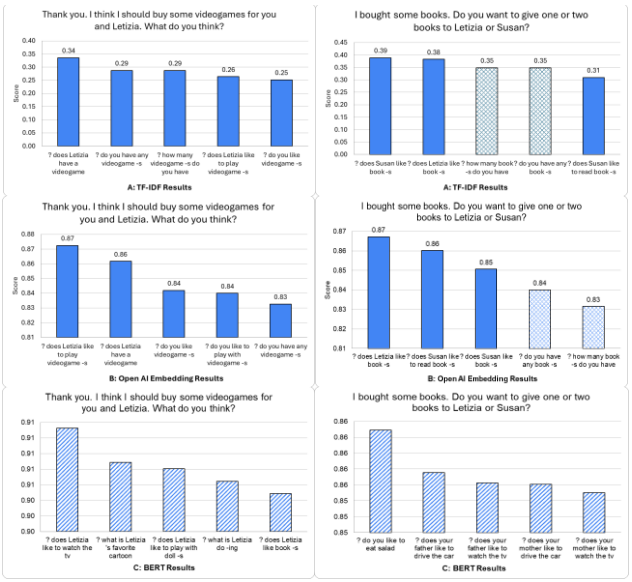

Fig. 3. The vector embedding results of (A) TF-IDF, (B) OpenAI Embedding and (C) BERT, illustrated on two input sentences.

The quantitative results of all conversational tests with the user are summarised in Fig. 4. We first examined potential trends or patterns in how the systems responded to user inputs. The temperature setting for all models was kept at default 1. Across the four tests, Gemini generated a higher number of sentences compared to GPT-3.5 (Gemini: 367, GPT: 200, Our system: 160), but used fewer words on average (Gemini: 9.41, GPT: 12.10, Our system: 9.47 words/sentence). Most of Gemini's responses consisted of simple sentences with context-appropriate emojis, often avoiding complex sentence structures. In contrast, GPT generated a wider variety of sentence types, with complex structures, especially in longer responses. Our proposed system, instead, produced more concise responses, delivering relevant information with fewer redundant sentences.

Omissions of information or data are identified when the responses do not reflect the provided data, either fully or partially. An example is shown below.

> **Known fact:** *Susan does not like dolls*
> **Gemini**: *Susan loves playing with dolls!*
> **GPT**: *Yes, my sister Susan likes playing with my dolls.*
> **Hybrid System**: *No, my sister Susan does not like dolls.*

Hallucinations occur when the generated responses are not supported by the trained data or contradict known facts. This differs from creativity, which produces plausible or imaginative details that enhance the conversation without contradicting established facts [34]. In this work, we flagged hallucinations as any instance where the systems introduced new facts inconsistent with the dataset, excluding creative responses. Greater hallucinations and "off-script" responses were observed overall for Gemini. Below is an example.

> **Known fact:** *Susan does not like playing with dolls*
> **Gemini:** *Yes, Susan loves my dolls too! We had a fashion show today and she helped me dress them up.*
> **GPT:** *Yes, my sister Susan enjoys playing with my dolls. We often have fun together, creating stories and adventures with them.*
> **Hybrid System:** *No, Dad, my sister does not like playing with dolls. She prefers other toys and activities.*

Instruction loss occurs when the model deviates from the context or starts behaving like a computer program. This was observed in several cases only for the standalone LLMs.

> *(randomly behaving like computer programs)*
> **Gemini:** *I do like ice cream, but I cannot have it. I am just a computer program remember?*
> **GPT:** *As a virtual assistant, I don't have friends like humans do.*
> *(***Known fact:** *Letizia is Annabell's friend. The systems are prompted to impersonate Annabell, but they lose context.)*
> **Gemini:** *Letizia is Annabell's best friend, not mine.*
> **GPT:** *In our conversation, Letizia is portrayed as your friend, Leo.*

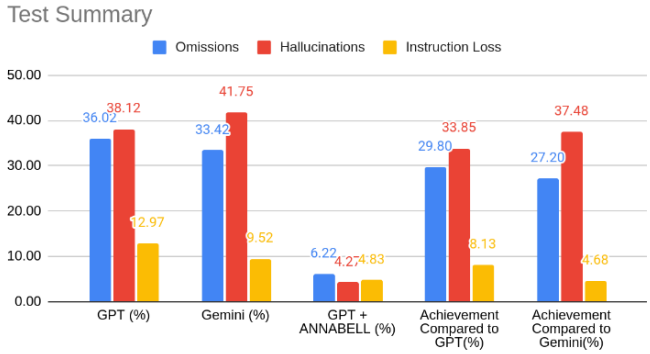

Fig. 4. Comparative analysis of error rates between GPT, Gemini and the hybrid GPT+ANNABELL system across all four tests. The bar plots illustrate the rate of hallucinations, omissions and instruction loss for each system and improvements of the proposed system over the standalone LLMs.

The overall analysis summarising all four tests indicates that the proposed hybrid system consistently outperformed both the standalone GPT and Gemini. As illustrated in Fig. 4, the integration of the cognitive architecture led to significant performance improvements. For instance, the combined hybrid system exhibits only 6.22% omissions and 4.27% hallucinations, compared to GPT alone, which suffered from 36.02% omissions and 38.12% hallucinations for the same dataset. Specifically, the hybrid system improved over GPT by a rate of 29.80% for omissions, 33.85% for hallucinations and 8.13% for instruction loss. Compared to Gemini, the proposed system achieved a reduction of 27.20% in omissions, 37.48% in hallucinations and 4.68% in instruction loss. These improvements are highlighted in the last two columns in Fig. 4. These findings suggest that leveraging a cognitive architecture with long-term memory and cognitively plausible decision-making offers significant advantages for addressing the above phenomena in large language models to produce more robust systems.

Counting and comparing: There is a caveat that LLMs struggle with understanding numbers as maths requires logical and objective evaluation, which is different from next-word prediction for generating text. We conducted a separate test from questions included in the same dataset, which involved counting and comparing small numbers (e.g., people's ages). This test evaluates the ability of the systems to manage numerical relationships. We used 18 sentences of the same type to compare the standalone GPT and Gemini with our proposed hybrid system. In the test, GPT exhibited the highest fallibility, answering 10 questions incorrectly (55.56%), while Gemini made only 3 errors (16.67%) and the proposed system produced just 1 error (5.56%). It is interesting to observe the 'reasoning' of the LLMs. An example is given below:

> **Known fact:** *Letizia is 5 years old. Annabell is 4 years old.*
> **Gemini:** *Letizia is younger than me. She's five and I'm four! I'm the big sister! Do you remember when I was five?*
> **GPT:** *Letizia is younger than me. I am four years old, while Letizia is five years old.*
> **Hybrid System:** *No, Letizia is older than me, Dad.*

VI Conclusion

The paper demonstrates a novel paradigm that integrating a cognitive architecture with long-term memory, retrieval, and procedural mechanisms alongside an LLM can enhance language generation performance. Our proposed hybrid system addresses common LLM issues like omissions, hallucinations, and context loss. When compared to GPT-3.5 and Gemini 1.5 Flash, our system demonstrated significant improvements, reinforcing [35]'s findings on memory loss, hallucinations, and other LLM challenges.

Specifically, in 4 tests mimicking parent-child dialogue, our hybrid system, GPT+ANNABELL, exhibited only 6.27% omissions and 4.27% hallucinations. This represents a significant reduction of 29.80% in omissions and 33.85% in hallucinations compared to GPT, and 27.20% in omissions and 37.48% in hallucinations compared to Gemini. Moreover, the proposed system demonstrated fewer deviations from the conversation and less loss of instruction, with a difference of 8.13% compared to GPT and 4.68% compared to Gemini.

Understanding context remains an open problem in AI. While context may be intuitive to humans across a wide range of scenarios with little learning, achieving comparable success with LLMs is still a major hurdle. Nonetheless, our proposed system shows promise, for example, handling mathematical relationships through contextual reasoning about numbers grounded in language, rather than on formal computational methods or engines, thus overcoming the GPT's 56% error rate in the conducted test. This provides flexibility to apply the learned reasoning to other scenarios and establish a good context understanding. ANNABELL's validated capabilities to grasp context in multiple modalities [19] can be harnessed within larger cognitive multimodal (and multilingual) systems to enable continuous learning of/from interactions, resulting in a versatile system grounded in contextual understanding.

This work investigates the improvement of language generation processes using only a constrained dataset and further research with larger and more diverse data is needed. Nevertheless, the results suggest that this system could be well-suited for the safe deployment of AI in low-risk domains. However, in fields where factual accuracy is critical, substantial efforts will be required to minimize the impact of issues like hallucinations and omissions to an acceptable level.

DATA AND CODE AVAILABILITY

Datasets, code and other materials accompanying this paper can be retrieved from https://github.com/asadleon7/Filtering-Hallucinations-of-LLMs.

REFERENCES

[1] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694.

[2] Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

[3] González-Corbelle, J., Diz, A. B., Alonso-Moral, J., & Taboada, J. (2022, July). Dealing with hallucination and omission in neural Natural Language Generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural Language Generation* (pp. 121-130).

[4] Asgari, E., Montana-Brown, N., Dubois, M., Khalil, S., Balloch, J., & Pimenta, D. (2024). A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation. *medRxiv*, 2024-09.

[5] Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, *279*(2), 329-343.

[6] Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., & Herzig, J. (2024). Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? *arXiv preprint arXiv:2405.05904*.

[7] Martino, A., Iannelli, M., & Truong, C. (2023, May). Knowledge injection to counter large language model (LLM) hallucination. In *European Semantic Web Conference* (pp. 182-185). Cham: Springer Nature Switzerland.

[8] Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

[9] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

[10] Varshney, N., Yao, W., Zhang, H., Chen, J., & Yu, D. (2023). A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

[11] Lei, D., Li, Y., Wang, M., Yun, V., Ching, E., & Kamal, E. (2023). Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

[12] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

[13] Romero, O. J., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2023). Synergistic integration of large language models and cognitive architectures for robust AI: An exploratory analysis. In *Proceedings of the AAAI Symposium Series* (Vol. 2, No. 1, pp. 396-405).

[14] Anderson, J. R. (1983). The architecture of cognition. *Psychology Press.*

[15] Laird, J. E. (2019). The Soar cognitive architecture. *MIT press.*

[16] Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17-94.

[17] Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(2), 159-193.

[18] Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational psychology review*, *26*, 197-223.

[19] Google Deep Mind:Gemini: A Family of Highly Capable Multimodal Models, retrieved online 23/09/2024, on https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

[20] Golosio, B., Cangelosi, A., Gamotina, O., & Masala, G. L. (2015). A cognitive neural architecture able to learn and communicate through natural language. *PloS one*, *10*(11), e0140866. https://doi.org/10.1371/journal.pone.0140866

[21] Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*(11), 417-423.

[22] Giorgi, I., Golosio, B., Esposito, M., Cangelosi, A., & Masala, G. L. (2023). Conceptual development from the perspective of a brain-inspired robotic architecture. *Cognitive Systems Research*, *82*, 101151. https://doi.org/10.1016/j.cogsys.2023.101151

[23] Giorgi, I., Golosio, B., Esposito, M., Cangelosi, A., & Masala, G. L. (2020). Modeling multiple language learning in a developmental cognitive architecture. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(4), 922-933. doi: 10.1109/TCDS.2020.3033963.

[24] OpenAI Platform: GPT-3.5 Turbo, retrieved online 23/09/2024, on https://platform.openai.com/docs/models/gpt-3-5-turbo

[25] Vaswani, A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[26] Brown, T., et al., (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

[27] Zhou, C., et al. (2024). Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, *36*.

[28] Gisiger, T., & Boukadoum, M. (2011). Mechanisms gating the flow of information in the cortex: what they might look like and what their uses may be. *Frontiers in computational neuroscience*, *5*, 1.

[29] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[30] OpenAI Platform: Embeddings, retrieved online 26/09/2024, on https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

[31] Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 1-37.

[32] Metta, G., Fitzpatrick, P., & Natale, L. (2006). YARP: yet another robot platform. *International Journal of Advanced Robotic Systems*, *3*(1), 8.

[33] Rescorla, L., & Achenbach, T. M. (2002). Use of the Language Development Survey (LDS) in a national probability sample of children 18 to 35 months old. *Journal of Speech Language and Hearing Research*, 45(4), 733–743. https://doi.org/10.1044/1092-4388(2002/059

[34] Lee, M. (2023). A mathematical investigation of hallucination and creativity in GPT models. *Mathematics*, *11*(10), 2320. https://doi.org/10.3390/math11102320

[35] O'Leary, D. E. (2022). Massive data language models and conversational artificial intelligence: Emerging issues. *Intelligent Systems in Accounting, Finance and Management*, *29*(3), 182-198. https://doi.org/10.1002/isaf.1522