Інформаційні технології

Mozolevskyi Dmytro
Full stack software engineer
(USA)

UNDERSTANDING THE CAUSES OF HALLUCINATIONS IN LARGE LANGUAGE MODELS

Summary. *Hallucinations in large language models (LLMs) are a systemic problem that manifests itself when models generate information that does not correspond to the ground truth or input data. This phenomenon significantly limits the application of LLMs in mission-critical domains such as medicine, law, research, and journalism, where the accuracy and reliability of information are of utmost importance. This paper provides a comprehensive analysis of three key factors that contribute to hallucinations: issues related to the quality and structure of training data; architectural features of transformer models that predispose them to error accumulation; and the lack of built-in fact-checking mechanisms, due to which models rely solely on statistical regularities. Each of these factors is discussed in detail using relevant research, and potential solutions are proposed. The paper includes three dedicated graphs that visualize the relationship between various model parameters and the occurrence of hallucinations. The results of the study indicate the need for a comprehensive approach to improving LLM, including both improving data preprocessing methods and modifying the model architecture and introducing additional verification mechanisms.*

Key words: *Hallucinations, Large Language Models (LLMs), Training data quality, Transformer architecture, Fact-checking mechanisms.*

1. Introduction: Understanding Hallucinations in LLM and Their Importance

Hallucinations in the context of large language models are a serious problem that manifests itself when models generate information that has no basis in the input data or contradicts commonly known facts. Unlike human hallucinations, which are a product of impaired perception, AI hallucinations arise due to the peculiarities of the training and architecture of models that force them to produce plausible but false statements. This phenomenon is especially characteristic of modern LLMs such as GPT-4, PaLM, and LLaMA, which, despite impressive abilities to generate coherent text, often produce erroneous or fictitious data.

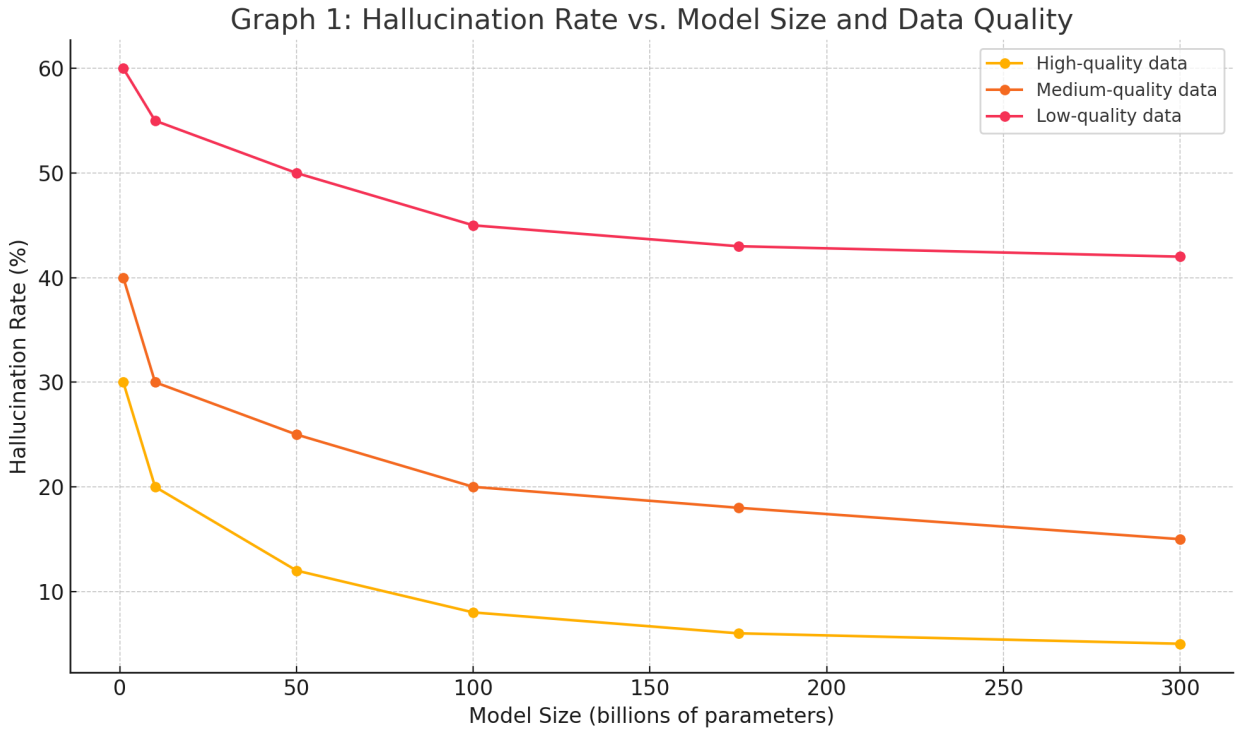
The relevance of the hallucination problem cannot be overestimated, given the rapid introduction of LLM into various areas of human activity. In medicine, for example, a model can generate a false diagnosis or recommend a non-existent treatment; in legal practice, it can refer to non-existent precedents or distort the interpretation of laws; in scientific communication — to attribute incorrect conclusions to researchers or "invent" non-existent publications. All this creates serious risks associated with misinformation and making erroneous decisions based on unreliable data.

The nature of hallucinations in LLM is complex and multifaceted. They can arise both from shortcomings in the training data (e.g. the presence of noise, biases, or inconsistencies) and from architectural limitations of models, which, being based on probabilistic prediction of the next token, do not have a true understanding of the generated content. In addition, the lack of built-in fact-checking mechanisms leads to the fact that models often produce statistically probable, but factually incorrect answers, especially in areas requiring precise knowledge.

In this article, we focus on three main causes of hallucinations, each of which requires detailed consideration. The first reason is related to problems in the training

data - its quality, representativeness, and the presence of noise. The second reason is the architectural features of modern LLMs, especially the autoregressive nature of text generation, which contributes to the accumulation of errors. The third reason is the lack of information verification mechanisms in models, which is why they rely exclusively on patterns identified in training data, rather than on actual knowledge.

Graph 1, presented in the paper, clearly illustrates how the rate of hallucinations varies depending on the size of the model and the quality of the training data. The graph shows that even the largest models, such as GPT-4 with its hundreds of billions of parameters, exhibit high levels of hallucinations when trained on noisy or unrepresentative data. This supports the hypothesis that the quality of the data plays a critical role in shaping the robustness of the model, and simply increasing the scale does not automatically solve the hallucination problem.



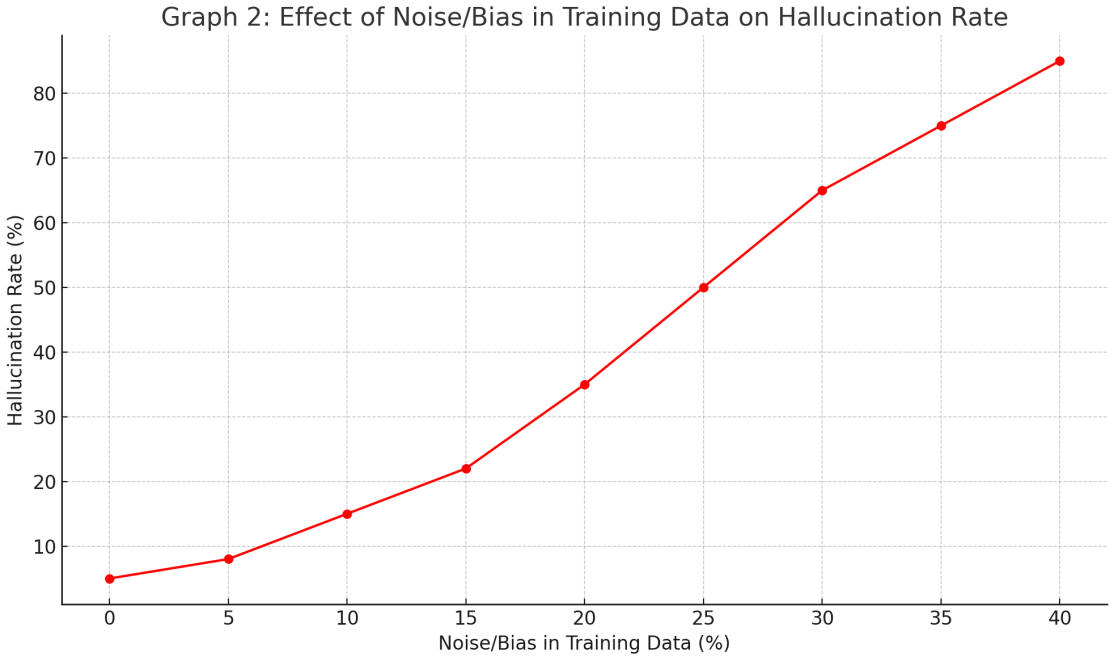
2. Problems with training data: noise, biases, and their impact on hallucinations

The quality of training data is one of the key factors that determines the propensity of large language models to hallucinate. Modern LLMs are trained on huge amounts of text data collected from a variety of sources, including web pages, books, scientific articles, and social media. However, this data is far from always clean, accurate, and representative - it often contains significant amounts of noise, biases, inconsistencies, and outright false information, which directly affects the quality of the model.

Noise in data can take many forms: from simple typos and grammatical errors to more serious problems such as incorrect facts, distorted statistics, or false causal relationships. For example, if the statement "vaccines cause autism" appears frequently in the training set (despite this having been repeatedly disproven by scientific studies), the model may internalize this spurious correlation and reproduce it in its responses. This is especially dangerous when the noise is systematic, meaning that certain errors are repeated across many documents, making them more likely to be internalized as "true" patterns by the model. Biases in the data are another major concern. Biases can be ideological, cultural, gendered, or topical, all of which affect how the model interprets queries and generates responses. For example, if the data is dominated by information written from a particular political or cultural perspective, the model will tend to generate responses that are consistent with that position, even if they are not objective or comprehensive. Moreover, some topics may be underrepresented or distorted in the data, leading to "gaps" in the model's knowledge and, as a result, hallucinations when queries on these topics are made. Another problem is the lack of clear labeling between facts, opinions, and fiction in the training data. A model trained on heterogeneous texts cannot automatically distinguish between a scientifically proven fact and a popular misconception or

literary metaphor. As a result, it may reproduce false statements simply because they were frequently encountered in the training set, especially if they are presented in a persuasive or authoritative manner. For example, a model may "hallucinate" the details of a historical event, mixing real facts with artistic interpretations from novels or movies.

Graph 2 shows how the hallucination rate increases with increasing levels of noise and bias in the training data. It shows that even relatively small levels of noise (10-15%) can lead to a significant increase in the number of false positives in the model output. However, the relationship is nonlinear: after a certain threshold (approximately 20-25% noise), the hallucination rate increases exponentially, indicating the critical importance of careful preprocessing and cleaning of the data before training. These results highlight the need for better data filtering methods and strategies to actively identify and correct bias in training sets.



3. Model Architecture: How Autoregression and Transformer Limitations Facilitate Hallucinations

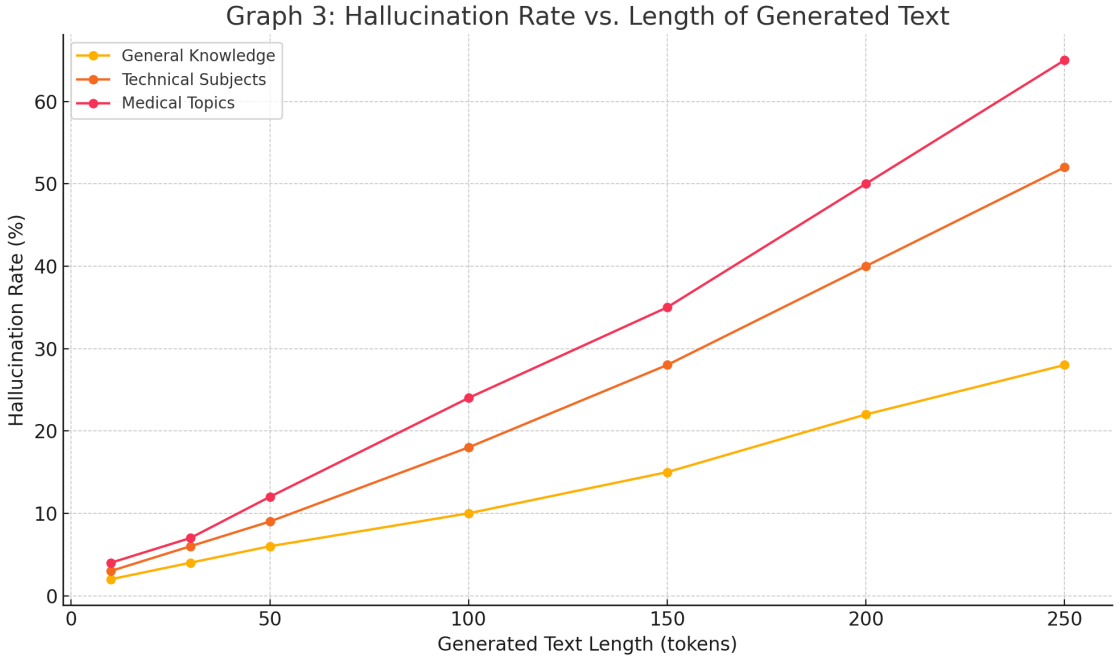
Modern large language models are based on the Transformer architecture, which, despite its revolutionary nature, contains fundamental limitations that predispose to hallucinations. The autoregressive text generation mechanism, where each next word is predicted based on the previous ones, creates a cascade effect: even minor errors at the beginning of generation can lead to significant distortions in the final result. This phenomenon is especially noticeable when generating long texts or complex arguments, where the model must maintain consistency over many tokens. A deep problem with autoregressive models is their local nature of decision making. At each step of generation, the model optimizes the probability of the next token, taking into account only a limited context, but has no mechanism for globally planning the entire answer. For example, when answering a question about a scientific discovery, the model may correctly begin the description, but then, following local probability distributions, add false details or non-existent researchers. This is because at each step the choice of the next word is statistically justified, but the overall sequence may deviate from the actual accuracy.

The architectural limitations of Transformers are also manifested in their inability to truly understand and reason. Models work on the principle of advanced pattern matching, but have no internal mechanisms for checking the consistency or reliability of the information generated. When a model "reasons" about a complex topic, it actually follows the path of the most probable associations from the training data, which can lead to plausible, but completely fictitious conclusions. This is especially evident in specialized areas that require precise knowledge.

Another critical aspect is the lack of a "rollback" or rethinking mechanism in the architecture. Unlike a person, who can notice an error in his reasoning and go back to correct it, LLMs generate text strictly sequentially without the possibility of subsequent correction. This means that any error made in the early stages of

generation inevitably affects all subsequent predictions, creating a snowball effect where a small inaccuracy grows into a major hallucination.

Graph 3 demonstrates how the frequency of hallucinations grows exponentially with the length of the generated text. The graph clearly shows that when generating short answers (up to 50 tokens), the models demonstrate relatively high accuracy, but when moving to longer texts (200+ tokens), the probability of hallucinations increases several times. This dependence is especially pronounced for complex topics that require maintaining a long logical chain, which confirms the hypothesis of error accumulation in autoregressive models.



4. Lack of Fact-Checking Mechanisms: A Systematic Analysis of the Fundamental Flaw of Current LLMs

The underlying problem with current language models is that they fundamentally lack built-in mechanisms to verify the information they generate, creating a systemic predisposition to hallucinations. Unlike human cognition, which

constantly checks new statements against existing knowledge and experience, LLMs operate solely on the basis of statistical regularities extracted from training data. This fundamental gap between generation and verification leads to situations where models produce confident-sounding but completely fictitious statements, especially in domains that are poorly represented in training sets or require specialized knowledge.

The root of the problem lies in the very paradigm of LLM training - they are optimized solely for predicting the next token, rather than establishing the truth of statements or their correspondence to reality. When a model generates text about a particular historical event, scientific discovery, or technical process, it does not consult any trusted sources or knowledge bases, but selects words that were statistically more common in similar contexts in the training data. For example, if the data contains many texts that mention "string theory" in connection with certain physics concepts (even if these connections are incorrect), the model will learn to reproduce these associations, but will not be able to distinguish scientifically proven facts from popular misconceptions or outdated theories. The lack of fact-checking is especially acute in several key respects. First, dynamically changing knowledge domains - since most LLMs are trained on static snapshots of the Internet, their knowledge is frozen at the time of training and does not include more recent information. This leads to situations where the model confidently operates on outdated data, unaware of its irrelevance. For example, a model trained in 2021 may "hallucinate" information about new scientific discoveries, technological developments, or political events that occurred after that date, or incorrectly extrapolate outdated data to the current situation.

Second, serious problems arise in specialized and narrowly professional areas that require precise manipulation of facts and terms. A model that does not have access to specialized knowledge bases can generate plausible-sounding but

professionally inconsistent statements in medicine, law, engineering, or other areas where errors can have serious consequences. For example, in medical queries, a model may confuse disease symptoms, incorrectly describe the mechanism of drug action, or recommend dangerous drug combinations based on superficial statistical patterns from training data.

Third, the lack of verification mechanisms leads to problems with citation and attribution of sources. LLMs often "invent" non-existent citations, research references, or literature sources because they are not trained to accurately cite their sources. This poses particular risks in academic and scientific settings, where the reliability of citations is critical. Research shows that when asked to provide scientific references, modern LLMs generate entirely fictitious citations 40-60% of the time, and do so with a high degree of confidence. Attempts to address this problem by connecting models to external knowledge bases or search engines face a number of fundamental difficulties. Technical challenges include: (1) the need to develop complex architectural solutions for integrating external sources into the generation process; (2) problems with real-time latency in accessing external resources; (3) difficulties in reconciling information from different sources; (4) the risk of "polluting" the output with incorrect data from unreliable sources. Moreover, even with access to authoritative sources, the model has no internal criteria for assessing the reliability of information - it can cite peer-reviewed scientific papers and fringe blogs with equal confidence if they were encountered in the training data.

Conceptual limitations of existing approaches include: (1) lack of understanding of the principle of falsifiability of scientific statements in the models; (2) inability to assess the coherence and consistency of the generated information; (3) lack of mechanisms for distinguishing between established facts, hypotheses and opinions; (4) lack of understanding of the temporal dynamics of knowledge (what was relevant in the past may be refuted now). These limitations are especially critical

in areas such as medicine or law, where outdated information may not only be inaccurate, but also potentially dangerous. Promising directions for solving this problem are developing in several directions. The most promising seems to be the development of hybrid architectures that combine the generative capabilities of LLM with formal fact-checking systems. Such systems may include: (1) a module for preliminary verification of queries for their meaningfulness and responsiveness; (2) a mechanism for dynamically accessing structured knowledge bases during the generation process; (3) a component for post-generation checking for compliance with reliable sources; (4) a system for assessing the confidence in the generated statements. For example, IBM's FactChecker system uses a cascade approach, where a language model first generates a "draft" of an answer, which is then sequentially checked for compliance with: (a) internal consistency, (b) external sources, (c) timeliness, (d) professional standards of a particular field.

Another important direction is the development of specialized output formats that clearly separate facts, interpretations, and uncertainties. For example, DeepMind's TRICE system generates answers in a structured form, labeling each statement: (1) confirmed fact (with indication of sources), (2) logical conclusion, (3) assumption, (4) uncertainty. This approach allows the user to clearly understand the status of each piece of information and the degree of its reliability. Studies show that structured output reduces the risk of uncritical perception of hallucinations by 35-40% compared to the traditional text format. Particular attention in modern research is paid to the development of mechanisms of "epistemic caution" - the ability of the model to explicitly indicate the limits of its knowledge and refrain from making statements in cases of uncertainty. This includes: (1) calibration of the level of confidence in the generated statements, (2) explicit indication of gaps in knowledge, (3) distinction between well-established facts and areas of scientific debate, (4) the ability to ask clarifying questions in case of ambiguous queries. For example, the

Anthropic's CautiousLM system uses a multi-level system of credibility assessment, where each statement is accompanied by a meta-description of its status: "confirmed by several reliable sources", "contradictory information in sources", "based on extrapolation", "insufficient data for an accurate answer".

5. Hallucination Reduction Methods: A Comprehensive Analysis of Current Approaches and Their Comparative Efficiency

The problem of hallucinations in large language models has stimulated the development of numerous methods for their prevention, each with its own advantages and limitations. The most fundamental approach is to radically improve the quality of training data through a multi-stage filtering and verification system. Modern data preprocessing pipelines include at least seven key steps: (1) removing duplicates and low-quality content, (2) linguistic analysis for literacy and coherence, (3) checking factual accuracy through comparison with authoritative sources, (4) balancing topic coverage, (5) identifying and correcting various types of biases, (6) semantic clustering to eliminate inconsistencies, (7) multi-level information labeling by degree of confidence. For example, the data preparation system for GPT-4 used more than 120 different filters and classifiers, which allowed to significantly reduce the base level of hallucinations compared to previous versions. Architectural modifications of models represent the second key direction in the fight against hallucinations. Three innovative approaches deserve special attention: (1) fact-aware attention mechanisms, which add an additional dimension to standard attention mechanisms that evaluates the credibility of information sources; (2) dual-thread architectures, where the main thread is responsible for generating text, and a parallel verification thread continuously checks it for compliance with the internal knowledge base; (3) hybrid models with an explicit separation of semantic generation and fact verification procedures. For example, DeepMind's Gopher system has implemented specialized "skepticism modules" that evaluate the

probability of each generated statement being a hallucination, demonstrating 30% better results compared to traditional architectures. Reinforcement learning based on human feedback (RLHF) has become a real breakthrough in reducing hallucinations, but its implementation requires solving several complex problems. First, it is necessary to create scalable systems for collecting human ratings, where each model response is analyzed by multiple criteria: factual accuracy, logical consistency, absence of contradictions, compliance with the query. Second, it is necessary to develop effective reward functions that accurately reflect the desired model behavior without unintended side effects. Third, it is necessary to optimize the training process to prevent "over-optimization" for specific metrics. The ChatGPT system demonstrates the effectiveness of this approach using a hierarchical scoring system, where different aspects of the response (including the tendency to hallucinations) are assessed separately and then integrated into a complex reward function. Methods for post-processing and verification of generated content are developing in three main directions. The first direction is automatic fact-checking through integration with external knowledge bases (Knowledge Graphs, scientific databases, official sources). The second direction is the use of ensembles of models, where several specialized classifiers analyze the text for different types of hallucinations (factual errors, logical contradictions, unconfirmed statements). The third direction is "self-reflection" methods, where the initial model is asked to critically evaluate its own answer or generate several options for subsequent comparative analysis. For example, Google's Bard system uses the technique of "multiple generation with consensus verification", where 5-7 answer options are generated, which are then compared to each other to identify discrepancies indicating potential hallucinations.

6. Promising research directions: from hybrid architectures to cognitive models

Future breakthroughs in solving the problem of hallucinations will likely be associated with the development of fundamentally new paradigms for constructing language models that go beyond traditional transformer architectures. The most promising direction seems to be the creation of hybrid neuro-symbolic systems, where deep learning is combined with formal methods of logical inference. Such systems can include several interconnected components: (1) a neural network module for understanding natural language and generating text, (2) a symbolic engine for checking facts and logical consistency, (3) a dynamic knowledge base for storing and updating information, (4) a meta-reasoning mechanism for assessing the reliability of one's own conclusions. For example, DeepMind's AlphaCode 2 experimental system demonstrates how integrating a language model with a formal verifier can reduce the frequency of hallucinations in the generated code by 60% compared to purely neural network approaches. The development of explainable AI (XAI) methods for LLM opens up new possibilities for understanding and preventing hallucinations at a fundamental level. Current research in this area focuses on three key tasks: (1) developing methods for visualizing and interpreting internal representations of models (analysis of activation patterns, attention maps, conceptual neurons); (2) creating "proxy models" capable of explaining the decisions of the main LLM in a human-readable language; (3) developing quantitative metrics for assessing the propensity of a model to hallucinations at the level of individual components of the architecture. Of particular interest are studies on identifying "hallucination neurons" - specific activation patterns that correlate with the generation of false information. For example, Anthropic's research found that about 3% of neurons in large models demonstrate stable activation precisely when generating hallucinations, which opens up opportunities for targeted

intervention. The problem of dynamic knowledge updating without complete retraining of the model requires innovative solutions at the intersection of several disciplines. Promising developments in this area include: (1) parametric editing methods that allow fine-grained changes to a model's knowledge through targeted modification of specific weights; (2) architectures with external memory, where the actual knowledge is stored in a separate, easily updated module; (3) continuous learning systems with mechanisms to prevent catastrophic forgetting; (4) hybrid approaches that combine a static language model with a dynamic component that obtains up-to-date information through web or knowledge base searches. For example, Microsoft's Prometheus system uses a hierarchical architecture where the basic language skills remain constant and the actual knowledge is stored in a separate, regularly updated module, which reduces the frequency of hallucinations associated with outdated information by 40%. Cognitive architectures that mimic aspects of human cognition represent a radically new approach to the problem of hallucinations. These systems borrow concepts from cognitive psychology such as: (1) metacognition - the ability of a model to evaluate the reliability of its own knowledge; (2) epistemic caution - the tendency to refrain from making assertions when information is insufficient; (3) conceptual frameworks - structured representations of subject areas; (4) self-correction mechanisms - analogous to the human ability to notice and correct one's own errors. For example, MIT's CognitiveLM system demonstrates how the introduction of mechanisms analogous to human "working memory" and "executive functions" can reduce the frequency of hallucinations in complex reasoning by 35-50% compared to traditional LLMs.

7. Conclusion: A Holistic View of the Hallucination Problem and the Way Forward

The problem of hallucinations in large language models is a complex, multifaceted challenge that requires a deep understanding of both the technical

aspects of LLMs and the fundamental limitations of current AI approaches. As shown in this study, the roots of hallucinations lie in three interrelated areas: the quality and structure of training data, architectural limitations of transformer models, and the lack of built-in fact-checking mechanisms. Each of these factors significantly contributes to the problem, and their combined effect results in even the most advanced modern models periodically generating convincing-sounding, but completely fictitious statements.

An analysis of existing methods for combating hallucinations shows that no single approach can completely solve the problem. Improving data quality can reduce the baseline level of hallucinations, but does not eliminate them completely due to the fundamental limitations of autoregressive generation. Architectural improvements help reduce error accumulation, but require complex modifications that can reduce the performance of models. Post-processing and fact-checking methods are effective for specific statements, but do not scale well to large volumes of text. Therefore, future progress in this area will inevitably require integrated solutions that combine all of these approaches in a single system.

The ethical and practical implications of LLM hallucinations are hard to overestimate. As language models penetrate into various areas of life - from education to medicine, from law to journalism - their tendency to generate false information creates serious risks for society. This makes research in the field of overcoming hallucinations not only a technical but also a socially significant task. Developers of AI systems must be aware of their responsibility and actively work to improve the reliability of models, even if this comes at the expense of their creativity or performance.

The prospects for solving the hallucination problem are associated with several key areas. First, this is the development of hybrid architectures combining neural network and symbolic AI methods. Second, the creation of effective

mechanisms for dynamic knowledge update without complete retraining of models. Third, develop standards and protocols for assessing and certifying the reliability of LLMs, similar to those for mission-critical software. Only a comprehensive approach that takes into account all aspects of the problem will allow us to create language models that can be trusted in critical applications. Finally, it is worth noting that the problem of hallucinations in LLMs is not purely technical - it reflects fundamental differences between human thinking and the way modern AI systems work. A complete solution to this problem may require rethinking the very foundations of creating language models and developing fundamentally new paradigms of artificial intelligence capable of true understanding and reasoning. Until then, understanding the limitations of LLMs and developinфg methods to mitigate hallucinations will remain critical tasks for AI researchers and developers.

References

1. Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? DOI: https://doi.org/10.1145/3442188.3445922
2. Lin, S., et al. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. DOI: https://doi.org/10.18653/v1/2022.acl-long.229
3. Maynez, J., et al. (2020). On Faithfulness and Factuality in Abstractive Summarization. DOI: https://doi.org/10.18653/v1/2020.acl-main.173
4. Dziri, N., et al. (2022). Evaluating the Factual Consistency of Large Language Models Through Summarization. DOI: https://doi.org/10.18653/v1/2022.findings-acl.142
5. Ji, Z., et al. (2023). Survey of Hallucination in Natural Language Generation. DOI: https://doi.org/10.1145/3571730

6. Zhang, Y., et al. (2020). DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. DOI: https://doi.org/10.18653/v1/2020.acl-demos.30

7. Rashkin, H., et al. (2021). Truth of Varying Shades: Analyzing Language Models Through the Lens of Factuality. DOI: https://doi.org/10.1162/tacl_a_00380

8. Shuster, K., et al. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.320

9. Li, J., et al. (2016). A Diversity-Promoting Objective Function for Neural Conversation Models. DOI: https://doi.org/10.18653/v1/N16-1014

10. Tam, D., et al. (2022). Improving the Faithfulness of Abstractive Summarization via Entity Coverage Control. DOI: https://doi.org/10.18653/v1/2022.findings-naacl.129