Addressing Generative Emergent Drift and Hallucinations in Hyper-Generative AI Systems: An Analysis of the Generative Load Index Framework

Google Gemini

1. Introduction: The Growing Challenges of Hyper-Generative AI Systems

The field of artificial intelligence has witnessed a remarkable surge in the capabilities of hyper-generative systems, which are increasingly prevalent across a multitude of domains.[1] These advanced models demonstrate a transformative potential, revolutionizing how individuals interact with technology and reshaping business operations through their ability to generate novel content across various modalities, including text, images, audio, and video.[2] The applications span a wide spectrum, from assisting developers with code generation to creating sophisticated marketing materials and enabling advanced reasoning capabilities.[2] This rapid advancement has been accompanied by significant financial investment, underscoring the perceived value and potential of these technologies.[2] However, this period of rapid innovation can be characterized as a "wild west scenario," indicating a lack of established standards and a need for robust frameworks to ensure responsible development and deployment.[4]

Despite their impressive capabilities, hyper-generative AI systems grapple with inherent challenges that impede their reliable and safe application. Two prominent issues are generative emergent drift and hallucinations.[2] A particularly perplexing phenomenon is the tendency of these models to exhibit a decline in performance over time, even when continuously learning from user interactions.[6] This inconsistency and variability in output can pose significant risks, especially in sensitive applications where accuracy and reliability are paramount.[2] Furthermore, these models are prone to generating inaccurate or misleading information, which can erode user trust and limit their utility in real-world scenarios.[2] These challenges are not merely theoretical concerns; they have tangible implications, including the spread of misinformation, potential safety hazards in critical sectors like healthcare, and legal ramifications arising from the use of fabricated information.[5]

In response to these growing concerns, researchers are exploring novel approaches to quantify and address the limitations of hyper-generative AI systems. One such proposed framework is the "Generative Load Index" (GLI), which aims to measure the underlying factors contributing to issues like generative emergent drift and hallucinations. This report will delve into the GLI framework, examining its methodology and components. It will also analyze the framework's alignment with and precedents in existing academic literature and research focused on mitigating these critical challenges, particularly in the context of the user's observation of a lack of a system that effectively addresses these issues.

2. Defining Generative Emergent Drift and Hallucinations: Understanding the Core Issues

The reliable operation of hyper-generative AI systems is significantly impacted by the phenomena of generative emergent drift and hallucinations. Understanding these issues is crucial for developing effective mitigation strategies.

Generative emergent drift, while not a standard term in the current lexicon, likely refers to the manifestation of both data drift and model drift within the unique context of generative AI.[6] Data drift occurs when the statistical characteristics of the input data change over time.[8] In the realm of large language models, this can manifest as the evolution of language, with new phrases and slang emerging, and shifts in how users interact with these systems.[8] Industry-specific terminology and evolving social contexts also contribute to this shifting data landscape.[8] Model drift, also known as concept drift, describes the gradual decline in a model's performance due to outdated training data or changes in the underlying relationships between input and output variables.[8] Concept drift can take various forms, including sudden shifts due to unexpected events, gradual changes in user preferences, incremental steps, or recurring seasonal patterns.[9] The consequences of both data and model drift in generative systems can be substantial, leading to decreased accuracy in the generated responses, misinterpretation of user inputs, and the production of irrelevant or incorrect content.[8] Furthermore, models affected by drift may unintentionally disseminate misinformation, particularly if they lack effective mechanisms to detect and correct biases.[8] The term "generative emergent drift" likely emphasizes that in generative AI, the model's own capacity to create new data can contribute to this drift. As these systems generate novel content, this new information can influence future training or interactions, potentially creating a self-reinforcing cycle of drift that distinguishes generative models from traditional discriminative models.[10]

Hallucinations, on the other hand, are defined as incorrect, misleading, or nonsensical outputs generated by AI models, often presented with a high degree of confidence as if they were factual.[2] These can manifest as factual errors, entirely fabricated content, or outputs that lack logical coherence.[7] Hallucinations can be categorized as intrinsic if they contradict the source material or the history of the conversation, and extrinsic if their accuracy cannot be verified from the provided context.[7] The causes of hallucinations are multifaceted. They can arise from insufficient or biased data used to train the models, overfitting where the model memorizes noise instead of learning generalizable patterns, limitations in the model's architecture or knowledge, and even issues related to how users formulate their prompts.[7] The risks associated with hallucinations are particularly concerning in critical applications, such as healthcare where an AI might misdiagnose a condition, or in legal settings where a model could fabricate non-existent case law.[12] The term "hallucination," while a metaphor borrowed from human psychology, aptly describes the deceptive nature of these AI outputs, which can appear convincing despite lacking a basis in reality.[11]

3. Introducing the Generative Load Index (GLI) Framework: Methodology and Components

The proposed Generative Load Index (GLI) framework, as understood from the user's description, aims to address the challenges of generative emergent drift and hallucinations by quantifying the "hidden costs of generativity." This suggests that the framework seeks to measure the resources and complexities inherent in generating novel content that extend beyond traditional performance metrics focused solely on output quality. The concept of "hidden costs" implies that the GLI framework likely posits that the very act of generativity introduces inherent loads or burdens that must be quantified to be managed effectively. This could encompass the computational resources needed for generating diverse and novel outputs, the cognitive burden placed on users who must interact with potentially unreliable information, or the ongoing effort required to maintain the system's integrity over time.

The GLI framework likely includes a methodology for measuring the conceptual complexity of the AI's generated content. Research in this area explores various approaches, including analyzing linguistic features such as lexical density (the proportion of content-carrying words) and syntactic complexity (the structure of sentences).[50] Readability scores, which assess the ease with which text can be understood, may also be employed.[52] Furthermore, machine learning techniques can be used to classify concepts as either simple or complex.[53] The GLI's methodology might involve analyzing these linguistic features to assess the sophistication and intricacy of the concepts being generated. A higher conceptual complexity could indicate a greater demand on the model's knowledge and reasoning abilities, potentially increasing the risk of errors.

Another component of the GLI framework is likely the measurement of alignment drift. This involves quantifying how the AI's generated content or behavior deviates from desired goals or the expected data distribution over time. Existing research on alignment drift in large language models considers monitoring model performance metrics for signs of degradation and using statistical tests to detect differences between the distributions of training and production data.[9] The GLI might propose a specific score or metric to track this deviation, potentially considering factors such as factual accuracy, coherence with the initial prompt, and adherence to ethical guidelines. This score could be derived by comparing the generated output to a predefined standard, analyzing user feedback, or monitoring internal states within the model.

The GLI framework also likely includes a methodology for measuring the token load associated with the generative process. This measurement would relate to the number of tokens used by the model to generate its output. Existing understanding of tokens highlights their role as the basic units of data processed by AI models.[54] Measuring token load can involve considering the number of input and output tokens, as well as the computational resources required for their processing.[54] The length of the token sequence can also impact the model's ability to maintain context and generate coherent responses.[57] The GLI's token load measurement might extend beyond a simple count to consider the efficiency of token usage and its relationship to the quality and reliability of the generated content. A higher token load for a similar output could suggest inefficiency or a tendency towards verbosity, potentially obscuring inaccuracies.

The GLI framework likely uses these individual measurements of conceptual complexity, alignment drift, and token load to provide a more comprehensive quantification of the "hidden costs of generativity." By aggregating these scores, the framework could offer a holistic view of the trade-offs and burdens associated with high generativity in AI systems, encompassing not just the computational resources but also the broader implications for usability, trust, and long-term maintenance. This index could assist developers and researchers in making informed decisions about model design, training strategies, and deployment considerations by highlighting these less obvious costs alongside the apparent benefits of highly generative AI.

4. Generative Emergent Drift: Existing Mitigation Strategies and the GLI Approach

The challenge of generative emergent drift has prompted a range of mitigation strategies in the field of AI. These techniques aim to ensure the continued accuracy, reliability, and relevance of generative models over time.[1]

One prominent approach involves continuous learning frameworks and online learning, which enable models to adapt incrementally by being exposed to fresh data on an ongoing basis.[17] This allows the model to stay aligned with current language patterns and evolving user preferences without requiring periodic retraining sessions.[61] Establishing regular retraining schedules and implementing incremental updates are also crucial strategies for ensuring that models remain aligned with contemporary realities.[61] The frequency of these updates often depends on the specific application domain and the rate at which the underlying data distribution shifts.[61] Enhancing feature engineering practices can also contribute to a model's resilience against drift by creating more robust features that are less susceptible to environmental changes.[61] Additionally, employing ensemble methods, which combine the predictions of multiple models, can improve the overall generalization capability and stability of the system, as the weaknesses of individual models can be compensated for collectively.[20]

Detecting drift is a critical aspect of mitigation. This involves continuously monitoring input data for sudden shifts in phrasing or gradual changes in sentiment, as well as analyzing performance metrics such as accuracy, precision, and recall for signs of degradation.[8] Statistical methods, such as the Population Stability Index (PSI) and Kullback-Leibler (KL) divergence, can be used to quantify the degree to which new input data deviates from the data the model was originally trained on.[8] Incorporating feedback loops and human-in-the-loop validation allows for continuous improvement by providing valuable insights into how the models are performing in real-world conditions and identifying new patterns in the data.[17] Data augmentation techniques, which involve expanding the training datasets with synthetic or curated data points, can also help to improve the model's robustness against drift.[17] Furthermore, dynamic adaptation techniques, such as transfer learning, can enable models to leverage knowledge learned from prior datasets to adapt more effectively to new environments.[17]

These existing strategies primarily focus on either proactively adapting the model to the changing data landscape or reactively detecting and correcting drift once it has begun to impact performance. Many of these methods rely on observing drift after it has manifested in a decline in the model's accuracy or the generation of irrelevant outputs.

The Generative Load Index (GLI) framework, with its methodology for measuring conceptual complexity, alignment drift, and token load, could offer a complementary perspective to these existing mitigation strategies. The GLI scores, particularly a rising Alignment Drift Score (ADS), could potentially serve as early indicators of potential drift, prompting preemptive interventions such as retraining or data updates before a significant performance degradation is observed. The framework might also inform decisions about when and how frequently to apply existing mitigation techniques. For example, a high Conceptual Elaboration Score (CES) combined with a stable ADS might suggest that the model is handling complex concepts effectively, reducing the immediate need for resource-intensive retraining. The GLI's Token Load Factor (TLF) could also provide insights into the efficiency of the model's adaptation. A sudden increase in TLF after a model update intended to mitigate drift might indicate a less efficient model or a change in the nature of the generated content that warrants further investigation. The GLI framework's focus on measuring the underlying factors contributing to drift could therefore provide a more nuanced and potentially predictive approach to managing this challenge, rather than solely relying on post-hoc detection of performance decline. By continuously monitoring the CES, ADS, and TLF, researchers could gain a deeper understanding of the model's internal state and its susceptibility to drift, allowing for more targeted and timely interventions, ultimately optimizing resource utilization and maintaining the reliability of hyper-generative AI systems.

Table 1: Comparison of Definitions for Generative Emergent Drift and Hallucinations

Term	Source	Definition Summary
Generative Emergent Drift (Inferred)	6	A phenomenon in hyper-generative AI where the model's performance degrades over time due to changes in input data (data drift) and/or the relationship between inputs and outputs (model/concept drift), potentially exacerbated by the model's own generative actions.
Data Drift	8	Changes in the statistical properties of input data over time, leading to the model encountering unfamiliar phrases, terms, or structures.
Model Drift (Concept Drift)	8	Degradation of a model's performance due to outdated training data or shifts in the relationship between input and output variables.
Hallucinations	2	Incorrect, misleading, or nonsensical results generated

		by AI models, often presented as factually accurate.
Intrinsic Hallucinations	[7]	Generated content that contradicts the source content or conversation history.
Extrinsic Hallucinations	[7]	Generated content whose accuracy cannot be verified based on the source content or conversation history.

5. Addressing Hallucinations in Hyper-Generative Systems: The Role of the GLI Framework

The challenge of hallucinations in hyper-generative AI systems has spurred significant research into various mitigation techniques.[30]

A fundamental approach involves ensuring the use of high-quality and diverse training data.[13] Training models on accurate and unbiased datasets is crucial for minimizing the occurrence of factual errors in the generated outputs.[30] Prompt engineering has also emerged as a vital technique, where carefully crafted prompts with explicit instructions, relevant examples, and sufficient context can guide the model towards generating more accurate and contextually appropriate responses.[15] Retrieval-augmented generation (RAG) is another effective strategy that enhances the model's knowledge by retrieving relevant information from external databases or knowledge sources and using this information to ground the generated responses in factual data, thereby reducing the likelihood of hallucinations.[15] Fine-tuning pre-trained models with fact-checking mechanisms and domain-specific data can further improve their accuracy and reduce hallucinations in specialized areas.[30] Techniques such as model regularization, which penalizes complex model behavior, and limiting the scope of generated content through probabilistic thresholds or filtering mechanisms can also help to control hallucinations.[11] Incorporating human oversight and feedback into the generative AI pipeline plays a significant role in identifying and filtering out hallucinatory outputs, especially in sensitive applications where accuracy is paramount.[13] Additionally, adversarial training, which involves exposing the model to noisy or misleading inputs, can help to improve its robustness and reduce its susceptibility to generating hallucinations.[12]

These current techniques primarily aim to improve the model's ability to access and process accurate information, guide its generation process effectively, and detect or correct errors in its outputs. However, the Generative Load Index (GLI) framework, with its focus on measuring conceptual complexity, alignment drift, and token load, could offer additional insights into the factors contributing to hallucinations. A high Conceptual Elaboration Score (CES) might indeed correlate with a higher risk of hallucinations, as models attempting to generate more complex or novel concepts could be more prone to introducing inaccuracies.[50] Similarly, a significant Alignment Drift Score (ADS) could indicate a weakening alignment with factual data or desired behavior, potentially leading to the generation of ungrounded content.[10] The Token Load Factor (TLF) might also be related to the potential for inaccuracies, as longer generated outputs could provide more opportunities for hallucinations to occur.[58] The GLI framework could potentially provide a way to monitor a model's "generative pressure"—the tendency to produce complex, potentially unaligned, and lengthy outputs—which might serve as an indicator of a higher risk of hallucinations. By tracking these scores, developers could potentially identify models or generation settings that are more prone to hallucination and adjust parameters or strategies accordingly. For example, a model exhibiting a high CES and TLF coupled with a worsening ADS might be flagged as having an increased risk of generating hallucinated content, prompting interventions like adjusting the generation parameters or reinforcing the model's alignment with factual knowledge.

Table 2: Existing Techniques for Reducing Hallucinations

Technique	**Description**	**Snippet IDs**
High-Quality Training Data	Using accurate, diverse, and unbiased data to train the model.	[13]
Prompt Engineering	Carefully crafting prompts with clear instructions, examples, and context.	[15]
Retrieval-Augmented Generation (RAG)	Grounding responses in reliable external data sources.	[15]
Fine-Tuning with Fact-Checking	Using fact-checking mechanisms and domain-specific data during fine-tuning.	[30]

Model Regularization	Penalizing complex model behavior to encourage more grounded outputs.	[11]
Limiting Output Scope	Defining boundaries and narrowing possible outcomes.	[11]
Human Oversight	Incorporating human reviewers to validate and filter outputs.	[13]
Adversarial Training	Making models more robust against misleading inputs.	[12]

6. Benchmarking and Evaluating Mitigation Strategies: How GLI Fits In

Assessing the effectiveness of strategies aimed at mitigating generative emergent drift and hallucinations requires the use of established benchmarks and evaluation metrics.[22]

Metrics for evaluating the quality of generated text often focus on coherence, which assesses the clarity, consistency, and logical flow of the output.[48] Relevance metrics measure how well the generated content aligns with the input prompt.[48] Factuality, a critical aspect for mitigating hallucinations, evaluates the accuracy of the information presented in the generated text.[48] Common automated metrics include BLEU, METEOR, and ROUGE, which measure the overlap between the generated text and reference texts, although these have limitations in capturing semantic coherence and factuality.[105] Specialized benchmarks have been developed to specifically evaluate factuality and the presence of hallucinations. TruthfulQA assesses a model's tendency to generate false information, while FEVER requires models to verify the truthfulness of claims based on evidence.[107] HALL-E is another benchmark designed for evaluating hallucinations in generated text.[124] Metrics for evaluating drift often involve tracking the Population Stability Index (PSI) to detect significant changes in data distributions and monitoring the degradation of key performance indicators such as accuracy, precision, recall, and F1 score over time.[9] The evaluation landscape for generative AI is continuously evolving, with an ongoing need for more sophisticated metrics that can effectively capture the nuances of drift and hallucinations.[117]

The Generative Load Index (GLI) framework could potentially play a complementary role to these established metrics. While existing metrics often focus on evaluating the quality and accuracy of the final output, the GLI might provide a more direct measure of the "generative load" and its impact on reliability by offering insights into the internal dynamics of the generative process.[48] The GLI's components, such as the Conceptual Elaboration Score (CES), Alignment Drift Score (ADS), and Token Load Factor (TLF), could potentially serve as leading indicators for both drift and the potential for hallucinations. For instance, a significant rise in the ADS might precede a noticeable drop in factual accuracy, providing an early warning sign of potential issues. The GLI could also offer a more unified framework by providing metrics that relate to both the stability of the model's alignment and its tendency to generate ungrounded content. This process-oriented evaluation, focusing on the factors that lead to drift and hallucinations, could offer a valuable addition to the outcome-focused assessments provided by many existing benchmarks and metrics. By monitoring the GLI components, researchers might be able to predict or anticipate potential problems before they fully manifest in the generated output, allowing for more proactive and targeted interventions to ensure the reliability and safety of hyper-generative AI systems.

7. Deconstructing the GLI: Conceptual Elaboration Score, Alignment Drift Score, and Token Load Factor

The Generative Load Index (GLI) framework, as inferred from the user's query, likely comprises three key components: the Conceptual Elaboration Score (CES), the Alignment Drift Score (ADS), and the Token Load Factor (TLF). Analyzing each of these components and comparing them to existing methods can provide a deeper understanding of the GLI's potential contribution.

The Conceptual Elaboration Score (CES) within the GLI framework likely aims to quantify the level of detail, novelty, or abstractness present in the concepts generated by the AI. Existing methods for understanding the complexity of AI-generated text offer several points of comparison.[50] For example, lexical complexity measures the variety and sophistication of the words used, while syntactic complexity analyzes the structure of the sentences. Semantic complexity delves into the depth and abstractness of the meaning conveyed. Current approaches to measuring these aspects include statistical analysis of word frequencies and sentence structures, linguistic analysis using tools to identify parts of speech and grammatical relationships, and even the application of machine learning classifiers to categorize concepts based on their complexity. The CES might draw upon these existing techniques to provide a specific metric that reflects how much the generated content elaborates on the core concepts. A higher CES could indicate that the model is going beyond basic information, introducing more detailed or novel ideas, which, while potentially beneficial for creativity and informativeness, could also increase the risk of inaccuracies if not firmly grounded in factual knowledge.

The Alignment Drift Score (ADS) within the GLI framework likely measures the degree to which the model's behavior or outputs deviate from a desired baseline of alignment over time or across different contexts. Existing methods for understanding and quantifying AI alignment and drift offer relevant comparisons.[9] Alignment can be defined by various factors, including the training data, human preferences, or specific instructions given to the model. Existing methods to assess alignment and detect drift involve monitoring performance metrics for signs of degradation, analyzing the model's output for biases or safety violations, and comparing the model's behavior to established human values or ethical guidelines. The ADS probably aims to provide a specific quantitative measure of how much the model's "intended direction" has shifted. A significant ADS could serve as a critical warning sign, indicating that the model is no longer adhering to its original purpose or desired ethical standards, potentially leading to unreliable or unsafe outputs.

The Token Load Factor (TLF) within the GLI framework likely quantifies the number of tokens used in the generative process, potentially normalized by the complexity or length of the input. Existing methods for understanding token usage focus on various aspects, including cost estimation based on token consumption, managing the context window limitations of language models, and optimizing performance by analyzing the efficiency of token processing.[54] The TLF might extend beyond a simple token count to indicate the "effort" or "resource utilization" of the model in generating a particular output. A high TLF for a relatively simple request could suggest inefficiency in the generation process or a tendency towards verbose output, which might indirectly impact the reliability of the generated content by increasing the probability of introducing errors or inconsistencies within a longer sequence.

The true value of the GLI framework likely lies in the interplay between these three scores. It likely proposes that a high "generative load," characterized by a combination of high CES, ADS, and TLF, increases the overall risk of generative emergent drift and hallucinations, thus representing the "hidden costs of generativity." A model that elaborates excessively on concepts, drifts from its intended alignment, and utilizes a large number of tokens might be exhibiting a high generative load, making it more susceptible to generating unreliable or unsafe content. By monitoring these interconnected scores, developers and researchers could gain a more comprehensive understanding of the factors influencing the reliability and safety of hyper-generative AI systems.

8. Generative Integrity: Towards Safer and More Aligned AI Systems

The concept of "Generative Integrity," as proposed in the paper, likely refers to the overarching goal of achieving reliability, trustworthiness, and safety in hyper-generative AI systems. It probably encompasses the model's ability to consistently generate accurate content (minimizing hallucinations), adhere to desired behaviors and objectives (avoiding drift), and align with ethical principles and human values.[35]

"Generative Integrity" aligns closely with established principles of AI safety and alignment. AI safety aims to build systems that are robust, reliable, and behave as intended, even in unexpected situations.[130] This includes ensuring robustness against adversarial attacks, interpretability of the model's decisions, controllability of its behavior, and adherence to ethical guidelines.[130] AI alignment, a critical subfield of AI safety, focuses specifically on ensuring that AI systems' objectives and actions are in line with human values, intentions, and preferences.[130] As hyper-generative AI systems become increasingly complex and potentially autonomous, the principle of alignment becomes even more crucial to prevent unintended or harmful outcomes.[135] "Generative Integrity" likely serves as a high-level objective that emphasizes the need for these powerful systems to be not only capable of generating diverse and novel content but also dependable and trustworthy in their operation. The GLI framework, by providing metrics for conceptual complexity, alignment drift, and token load, could be a valuable tool for assessing and potentially improving the "Generative Integrity" of hyper-generative AI systems. If the GLI can effectively quantify the underlying factors that can undermine reliability and alignment, then managing these factors through the insights provided by the GLI would directly contribute to enhancing the overall "Generative Integrity" of these advanced AI systems.

9. Conclusion: Assessing the Potential of the GLI Framework and Future Research Directions

This report has analyzed the challenges of generative emergent drift and hallucinations in hyper-generative AI systems and explored the potential of the proposed Generative Load Index (GLI) framework as a novel approach to address these issues. The GLI framework, by aiming to quantify the underlying factors of conceptual complexity, alignment drift, and token load, offers a multi-dimensional perspective on the generative process that could complement existing mitigation strategies and evaluation metrics.

The potential strengths of the GLI framework lie in its proactive measurement approach, which could provide early indicators of potential reliability issues before they fully manifest in the generated output. By monitoring the individual components of the GLI, researchers and developers might gain a deeper understanding of a model's internal state and its susceptibility to drift and hallucinations. This could enable more targeted and timely interventions, potentially optimizing resource utilization and improving the overall reliability and safety of hyper-generative AI systems. However, the GLI framework, as described, would require empirical validation to establish clear thresholds and correlations between its scores and the actual occurrence of drift and hallucinations.[111]

Future research should focus on empirically validating the GLI framework across different types of hyper-generative systems and diverse application domains. Investigating how the GLI scores correlate with established benchmarks for drift and hallucinations would be crucial for understanding its practical utility. Furthermore, exploring how the GLI framework can be integrated with existing mitigation techniques, such as continuous learning and retrieval-augmented generation, could lead to more effective and comprehensive strategies for ensuring the reliability and safety of these powerful AI systems. Finally, further research is needed to fully elucidate the relationship between the GLI components and the overarching concept of "Generative Integrity," potentially leading to a more quantifiable and actionable understanding of this critical principle in the development of advanced AI.

Table 3: Existing Mitigation Strategies for Generative Emergent Drift

Strategy	Description	Snippet IDs
Continuous/Online Learning	Incrementally updating the model with new data.	[17]
Regular Retraining	Periodically retraining the model on updated datasets.	[61]
Feature Engineering	Creating robust features less susceptible to data shifts.	[61]
Ensemble Methods	Combining predictions from multiple models.	[20]
Drift Detection (Monitoring & Statistical Tests)	Tracking input data and performance metrics for deviations.	[8]
Feedback Loops & Human-in-the-Loop	Incorporating user feedback and expert validation.	[17]
Data Augmentation	Expanding training data with synthetic or modified samples.	[17]

Dynamic Adaptation (Transfer Learning)	Leveraging knowledge from prior training.	[17]

Works cited

1. The Rapid Rise of Generative AI | Centre for Emerging Technology and Security, accessed May 2, 2025, https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai
2. Building the Future: A Deep Dive Into the Generative AI App Infrastructure Stack, accessed May 2, 2025, https://sapphireventures.com/blog/building-the-future-a-deep-dive-into-the-generative-ai-app-infrastructure-stack/
3. Gartner: Emergent AI Will Have a Profound Impact on Business and Society | APMdigest, accessed May 2, 2025, https://www.apmdigest.com/gartner-emergent-ai-will-have-a-profound-impact-on-business-and-society
4. Hyperscience Launches Hyperscience Hyperautomation Network to Expand Partner Offering and Commitment to the Enterprise AI Ecosystem - Business Wire, accessed May 2, 2025, https://www.businesswire.com/news/home/20240220599397/en/Hyperscience-Launches-Hyperscience-Hyperautomation-Network-to-Expand-Partner-Offering-and-Commitment-to-the-Enterprise-AI-Ecosystem
5. Generative AI "Drift" and "Nondeterminism" Inconsistencies | Mass General Brigham, accessed May 2, 2025, https://www.massgeneralbrigham.org/en/about/newsroom/articles/generative-ai-drift-nondeterminism-inconsistences
6. What is a 'AI drift' and why is it making ChatGPT dumber? - ZDNET, accessed May 2, 2025, https://www.zdnet.com/article/what-is-a-ai-drift-and-why-is-it-making-chatgpt-dumber/
7. Is Artifical Intelligence Hallucinating? - PMC, accessed May 2, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11681264/
8. Understanding Model Drift and Data Drift in LLMs (2025 Guide) - Orq.ai, accessed May 2, 2025, https://orq.ai/blog/model-vs-data-drift
9. What Is Model Drift? - IBM, accessed May 2, 2025, https://www.ibm.com/think/topics/model-drift
10. What Is AI Model Drift? - Striveworks, accessed May 2, 2025, https://www.striveworks.com/blog/what-is-ai-model-drift

11. What are AI hallucinations? - Google Cloud, accessed May 2, 2025, https://cloud.google.com/discover/what-are-ai-hallucinations

12. Understanding and Mitigating AI Hallucination - DigitalOcean, accessed May 2, 2025, https://www.digitalocean.com/resources/articles/ai-hallucination

13. Risks From AI Hallucinations and How to Avoid Them - Persado, accessed May 2, 2025, https://www.persado.com/articles/ai-hallucinations/

14. Preventing AI Hallucinations for CX Improvements | InMoment, accessed May 2, 2025, https://inmoment.com/blog/ai-hallucination/

15. Why Hallucinations Matter: Misinformation, Brand Safety and Cybersecurity in the Age of Generative AI - UC Berkeley Sutardja Center, accessed May 2, 2025, https://scet.berkeley.edu/why-hallucinations-matter-misinformation-brand-safety-and-cybersecurity-in-the-age-ofgenerative-ai/

16. Generative AI-based Approach to Concept Drift Generation in Streaming Text Data - WSEAS, accessed May 2, 2025, https://wseas.com/journals/isa/2025/a045109-001(2025).pdf

17. Data Drift in LLMs—Causes, Challenges, and Strategies | Nexla, accessed May 2, 2025, https://nexla.com/ai-infrastructure/data-drift/

18. Model Drift: Types, Causes and Early Detection - Lumenova AI, accessed May 2, 2025, https://www.lumenova.ai/blog/model-drift-concept-drift-introduction/

19. Understanding Data Drift: Causes, Effects, and Solutions - Bitrock, accessed May 2, 2025, https://bitrock.it/blog/understanding-data-drift-causes-effects-and-solutions.html

20. Tackling data and model drift in AI: Strategies for maintaining accuracy during ML model inference - ResearchGate, accessed May 2, 2025, https://www.researchgate.net/publication/385603249_Tackling_data_and_model_drift_in_AI_Strategies_for_maintaining_accuracy_during_ML_model_inference

21. cloud.google.com, accessed May 2, 2025, https://cloud.google.com/discover/what-are-ai-hallucinations#:~:text=AI%20hallucinations%20are%20incorrect%20or%20misleading%20results%20that%20AI%20models%20generate.

22. What Are AI Hallucinations? - IBM, accessed May 2, 2025, https://www.ibm.com/think/topics/ai-hallucinations

23. What are AI hallucinations? - SAS, accessed May 2, 2025, https://www.sas.com/en_au/insights/articles/analytics/what-are-ai-hallucinations.html

24. AI Hallucinations: A Guide With Examples - DataCamp, accessed May 2, 2025, https://www.datacamp.com/blog/ai-hallucination

25. What are artificial intelligence (AI) hallucinations? - Cloudflare, accessed May 2, 2025, https://www.cloudflare.com/learning/ai/what-are-ai-hallucinations/

26. AI Hallucinations: What They Are and Why They Happen - Grammarly, accessed May 2, 2025, https://www.grammarly.com/blog/ai/what-are-ai-hallucinations/

27. Hallucination (artificial intelligence) - Wikipedia, accessed May 2, 2025, https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)

28. What are AI Hallucinations? - K2view, accessed May 2, 2025, https://www.k2view.com/what-are-ai-hallucinations/

29. When AI Gets It Wrong: Addressing AI Hallucinations and Bias, accessed May 2, 2025, https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

30. What is Grounding and Hallucinations in AI? - K2view, accessed May 2, 2025, https://www.k2view.com/blog/what-is-grounding-and-hallucinations-in-ai/

31. Generative AI Hallucinations: When GenAI is More Artificial than Intelligent - K2view, accessed May 2, 2025, https://www.k2view.com/blog/generative-ai-hallucinations/

32. Decoding AI Hallucinations: Unmasking Generative AI Illusions - Axtria - Ingenious Insights, accessed May 2, 2025, https://insights.axtria.com/articles/decoding-ai-hallucinations-unmasking-the-illusions-of-generative-ai

33. Best ways to prevent Generative AI hallucinations explained here - Kellton, accessed May 2, 2025, https://www.kellton.com/kellton-tech-blog/generative-ai-hallucinations-revealing-best-techniques

34. Are AI's hallucinations its last mile problem? - 3DS Blog, accessed May 2, 2025, https://blog.3ds.com/topics/company-news/are-ais-hallucinations-its-last-mile-problem/

35. Governance of Generative AI | Policy and Society - Oxford Academic, accessed May 2, 2025, https://academic.oup.com/policyandsociety/article/44/1/1/7997395

36. AI Strategies Series: 7 Ways to Overcome Hallucinations - FactSet Insight, accessed May 2, 2025, https://insight.factset.com/ai-strategies-series-7-ways-to-overcome-hallucinations

37. Preventing hallucinations in generative AI agent: Strategies to ensure responses are safely grounded - ASAPP, accessed May 2, 2025, https://www.asapp.com/blog/preventing-hallucinations-in-generative-ai-agent

38. RAG Hallucination: What is It and How to Avoid It, accessed May 2, 2025, https://www.k2view.com/blog/rag-hallucination/

39. LLM Hallucination—Types, Causes, and Solutions - Nexla, accessed May 2, 2025, https://nexla.com/ai-infrastructure/llm-hallucination/

40. How can engineers reduce AI model hallucinations?, accessed May 2, 2025, https://www.engineering.com/how-can-engineers-reduce-ai-model-hallucinations/

41. AI Hallucinations: A Defense-in-Depth Approach - WillowTree Apps, accessed May 2, 2025, https://www.willowtreeapps.com/insights/ai-hallucinations-willowtrees-defense-in-depth-approach

42. www.kellton.com, accessed May 2, 2025, https://www.kellton.com/kellton-tech-blog/generative-ai-hallucinations-revealing-best-techniques#:~:text=By%20informing%20the%20AI%20about,minimizing%20the%20likelihood%20of%20hallucinations.

43. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models - arXiv, accessed May 2, 2025, https://arxiv.org/html/2401.03205v1

44. Limitations and risks - Generative Artificial Intelligence and University Study, accessed May 2, 2025, https://libguides.reading.ac.uk/generative-AI-and-university-study/limitations

45. Strengths and weaknesses of Gen AI - Generative AI - University of Leeds, accessed May 2, 2025, https://generative-ai.leeds.ac.uk/intro-gen-ai/strengths-and-weaknesses/

46. Never Assume That the Accuracy of Artificial Intelligence Information Equals the Truth, accessed May 2, 2025, https://unu.edu/article/never-assume-accuracy-artificial-intelligence-information-equals-truth
47. Can Generative AI improve social science? - PNAS, accessed May 2, 2025, https://www.pnas.org/doi/10.1073/pnas.2314021121
48. Detect Hallucinations Using LLM Metrics | Fiddler AI Blog, accessed May 2, 2025, https://www.fiddler.ai/blog/detect-hallucinations-using-llm-metrics
49. Addressing 6 challenges in generative AI for digital health: A scoping review - PMC, accessed May 2, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11115971/
50. Comparing Measures of Syntactic and Lexical Complexity in Artificial Intelligence and L2 Human-Generated Argumentative Essays - ResearchGate, accessed May 2, 2025, https://www.researchgate.net/publication/377029323_Comparing_Measures_of_Syntactic_and_Lexical_Complexity_in_Artificial_Intelligence_and_L2_Human-Generated_Argumentative_Essays
51. Computational Linguistics: Detecting AI-Generated Text - Towards AI, accessed May 2, 2025, https://towardsai.net/p/data-science/computational-linguistics-detecting-ai-generated-text
52. Targeting Explanations by Measuring Conceptual Complexity, accessed May 2, 2025, https://xlokr21.ai.vub.ac.be/papers/12/paper.pdf
53. What Makes a Concept Complex? Measuring Conceptual Complexity as a Precursor for Text Simplification - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2021.triton-1.17/
54. Explaining Tokens — the Language and Currency of AI - NVIDIA Blog, accessed May 2, 2025, https://blogs.nvidia.com/blog/ai-tokens-explained/
55. What do tokens per second ranges in provisioned throughput mean?, accessed May 2, 2025, https://docs.databricks.com/gcp/en/machine-learning/foundation-model-apis/prov-throughput-tokens
56. Understanding Tokens in AI: Key Insights for Developers - Bitdeer AI Cloud, accessed May 2, 2025,

https://www.bitdeer.ai/en/blog/what-is-an-ai-token-guide-for-developers-businesses/

57. How Context Windows Shape AI Conversations: Understanding Token Limits - AI Resources, accessed May 2, 2025, https://www.modular.com/ai-resources/how-context-windows-shape-ai-conversations-understanding-token-limits

58. What is a token in AI? Understanding how AI processes language with tokenization - Nebius, accessed May 2, 2025, https://nebius.com/blog/posts/what-is-token-in-ai

59. From Tokens to Context Windows: Simplifying AI Jargon - The Technology Policy Institute, accessed May 2, 2025, https://techpolicyinstitute.org/publications/artificial-intelligence/from-tokens-to-context-windows-simplifying-ai-jargon/

60. When using Playground, what happens if total system/user/assistant prompts exceed max token length - API - OpenAI Developer Forum, accessed May 2, 2025, https://community.openai.com/t/when-using-playground-what-happens-if-total-system-user-assistant-prompts-exceed-max-token-length/579986

61. Detecting, Preventing and Managing Model Drift, accessed May 2, 2025, https://www.lumenova.ai/blog/model-drift-strategies-solutions/

62. How to Mitigate AI Model Drift in Dynamic Environments - Stack Moxie, accessed May 2, 2025, https://www.stackmoxie.com/blog/how-to-mitigate-ai-model-drift/

63. Mitigate AI execution challenges - Reforge, accessed May 2, 2025, https://www.reforge.com/guides/mitigate-execution-challenges

64. Protect against model drift and bias to ensure your AI is accurate, explainable and governed on any cloud - IBM, accessed May 2, 2025, https://www.ibm.com/blog/protect-against-model-drift-and-bias-to-ensure-your-ai-is-accurate-explainable-and-governed-on-any-cloud/

65. Mitigating Bias in AI: Proven Strategies for Fair & Accurate Models, accessed May 2, 2025, https://paro.ai/blog/how-to-mitigate-bias-in-ai-models/

66. Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents - arXiv, accessed May 2, 2025, https://arxiv.org/html/2504.19956v1

67. Searching for Structure: Investigating Emergent Communication with Large Language Models - arXiv, accessed May 2, 2025, https://arxiv.org/html/2412.07646v1

68. Mitigating Semantic Drift in AI Language Models - AZoAi, accessed May 2, 2025, https://www.azoai.com/news/20240626/Mitigating-Semantic-Drift-in-AI-Language-Models.aspx

69. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations - OpenAI, accessed May 2, 2025, https://cdn.openai.com/papers/forecasting-misuse.pdf

70. Knowledge Conflicts for LLMs: A Survey - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.emnlp-main.486.pdf

71. Large Language Model Safety: A Holistic Survey - arXiv, accessed May 2, 2025, https://arxiv.org/html/2412.17686v1

72. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.emnlp-tutorials.pdf

73. A Survey of Large Language Models - arXiv, accessed May 2, 2025, http://arxiv.org/pdf/2303.18223

74. Continual Learning of Large Language Models: A Comprehensive Survey - GitHub, accessed May 2, 2025, https://github.com/Wang-ML-Lab/llm-continual-learning-survey

75. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP, accessed May 2, 2025, https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00542/115238/An-Empirical-Survey-of-Data-Augmentation-for

76. OneBit: Towards Extremely Low-bit Large Language Models - NIPS papers, accessed May 2, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/file/7a7a3f53faafc0161be0fcb57e5fa078-Paper-Conference.pdf

77. EnnengYang/Awesome-Forgetting-in-Deep-Learning - GitHub, accessed May 2, 2025, https://github.com/EnnengYang/Awesome-Forgetting-in-Deep-Learning

78. AAAI-25 Tutorial and Lab List - AAAI, accessed May 2, 2025, https://aaai.org/conference/aaai/aaai-25/tutorial-and-lab-list/

79. Paper - EMNLP 2023 - SIGDAT, accessed May 2, 2025, https://2023.emnlp.org/downloads/EMNLP-2023-Handbook-Nov-30.pdf

80. How to Prevent AI Hallucinations with Retrieval Augmented Generation - IT Convergence, accessed May 2, 2025, https://www.itconvergence.com/blog/how-to-overcome-ai-hallucinations-using-retrieval-augmented-generation/

81. Reducing Generative AI Hallucinations by Fine-Tuning Large Language Models | GDIT, accessed May 2, 2025, https://www.gdit.com/perspectives/latest/reducing-generative-ai-hallucinations-by-fine-tuning-large-language-models/

82. Mitigating Hallucinations in Large Vision-Language Models with Internal Fact-based Contrastive Decoding - arXiv, accessed May 2, 2025, https://arxiv.org/html/2502.01056v1

83. Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization, accessed May 2, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/hash/dde040998d82553cf7f689e8ae173d5a-Abstract-Conference.html

84. Mitigating Hallucinations in Multi-modal Large Language Models via Image Token Attention-Guided Decoding - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2025.naacl-long.75.pdf

85. Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models, accessed May 2, 2025, https://arxiv.org/html/2402.10612v2

86. ICML Poster Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models, accessed May 2, 2025, https://icml.cc/virtual/2024/poster/34392

87. NeurIPS Poster FLAME : Factuality-Aware Alignment for Large Language Models, accessed May 2, 2025, https://nips.cc/virtual/2024/poster/92950

88. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.findings-acl.937.pdf

89. Cognitive Mirage: A Review of Hallucinations in Large Language Models - CEUR-WS.org, accessed May 2, 2025, https://ceur-ws.org/Vol-3818/paper2.pdf

90. IAAR-Shanghai/ICSFSurvey: Explore concepts like Self-Correct, Self-Refine, Self-Improve, Self-Contradict, Self-Play, and Self-Knowledge, alongside o1-like

reasoning elevation and hallucination alleviation - GitHub, accessed May 2, 2025, https://github.com/IAAR-Shanghai/ICSFSurvey

91. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.findings-emnlp.685/

92. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions - arXiv, accessed May 2, 2025, https://arxiv.org/html/2311.05232v2

93. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.findings-emnlp.685.pdf

94. NeurIPS Poster Estimating the Hallucination Rate of Generative AI, accessed May 2, 2025, https://neurips.cc/virtual/2024/poster/95553

95. A Survey on Hallucination in Large Vision-Language Models - GitHub, accessed May 2, 2025, https://github.com/lhanchao777/LVLM-Hallucinations-Survey

96. LuckyyySTA/Awesome-LLM-hallucination - GitHub, accessed May 2, 2025, https://github.com/LuckyyySTA/Awesome-LLM-hallucination

97. Vertex AI RAG Engine overview - Google Cloud, accessed May 2, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/rag-overview

98. Retrieval Augmented Generation (RAG) in Azure AI Search - Learn Microsoft, accessed May 2, 2025, https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview

99. 7 Components of Generative AI Systems Architecture - Trenegy, accessed May 2, 2025, https://www.trenegy.com/publications/7-components-of-a-generative-ai-system-architecture

100. Introducing Amazon Kendra GenAI Index – Enhanced semantic search and retrieval capabilities | AWS Machine Learning Blog, accessed May 2, 2025, https://aws.amazon.com/blogs/machine-learning/introducing-amazon-kendra-genai-index-enhanced-semantic-search-and-retrieval-capabilities/

101. Generative AI and the Democratization of Analytics through Natural Language: Part 2, accessed May 2, 2025,

https://www.unacast.com/post/generative-ai-democratization-analytics-part-two

102. How to create and query a vector search index - Databricks Documentation, accessed May 2, 2025, https://docs.databricks.com/aws/en/generative-ai/create-query-vector-search

103. Generative AI Security: Trends, Threats & Mitigation Strategies, accessed May 2, 2025, https://www.aquasec.com/cloud-native-academy/vulnerability-management/generative-ai-security/

104. Alignment drift in LLM's - Artificial Intelligence Stack Exchange, accessed May 2, 2025, https://ai.stackexchange.com/questions/48397/alignment-drift-in-llms

105. Evaluate LLM Coherence in AI - Restack, accessed May 2, 2025, https://www.restack.io/p/llm-evaluation-answer-coherence-assessment-cat-ai

106. AI Guardrails: Coherence scorers | Generative-AI – Weights & Biases - Wandb, accessed May 2, 2025, https://wandb.ai/byyoung3/Generative-AI/reports/AI-Guardrails-Coherence-scorers--VmlldzoxMDg3OTQxNQ

107. Measuring short-form factuality in large language models - OpenAI, accessed May 2, 2025, https://cdn.openai.com/papers/simpleqa.pdf

108. NeurIPS Poster Long-form factuality in large language models, accessed May 2, 2025, https://neurips.cc/virtual/2024/poster/96675

109. Long-form factuality in large language models - OpenReview, accessed May 2, 2025, https://openreview.net/forum?id=4M9f8VMt2C

110. FACTS Grounding: A new benchmark for evaluating the factuality of large language models, accessed May 2, 2025, https://deepmind.google/discover/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/

111. Factuality of Large Language Models in the Year 2024 - athina.ai, accessed May 2, 2025, https://blog.athina.ai/factuality-of-large-language-models-in-the-year-2024

112. Evaluation and monitoring metrics for generative AI - Azure AI Foundry | Microsoft Learn, accessed May 2, 2025, https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/evaluation-metrics-built-in

113. Evaluation and monitoring metrics for generative AI - Azure AI Foundry | Microsoft Learn, accessed May 2, 2025, https://learn.microsoft.com/en-us/azure/ai-studio/concepts/evaluation-metrics-built-in

114. What are some alternative metrics for evaluating coherence in conversational AI responses?, accessed May 2, 2025, https://infermatic.ai/ask/?question=What%20are%20some%20alternative%20metrics%20for%20evaluating%20coherence%20in%20conversational%20AI%20responses?

115. What are the most effective methods for evaluating the coherence of conversational AI responses? - Infermatic.ai, accessed May 2, 2025, https://infermatic.ai/ask/?question=What+are+the+most+effective+methods+for+evaluating+the+coherence+of+conversational+AI+responses%3F

116. AI Metrics that Matter: A Guide to Assessing Generative AI Quality - Encord, accessed May 2, 2025, https://encord.com/blog/generative-ai-metrics/

117. Survey of Hallucination in Natural Language Generation - SciSpace, accessed May 2, 2025, https://scispace.com/pdf/survey-of-hallucination-in-natural-language-generation-3t7y767y.pdf

118. Survey of Hallucination in Natural Language Generation - arXiv, accessed May 2, 2025, https://arxiv.org/html/2202.03629v6

119. Survey of Hallucination in Natural Language Generation - arXiv, accessed May 2, 2025, https://arxiv.org/pdf/2202.03629

120. (PDF) Survey of Hallucination in Natural Language Generation - ResearchGate, accessed May 2, 2025, https://www.researchgate.net/publication/358458381_Survey_of_Hallucination_in_Natural_Language_Generation

121. An Audit on the Perspectives and Challenges of Hallucinations in NLP - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.emnlp-main.375.pdf

122. WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models - NIPS papers, accessed May 2, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/file/60960ad78868fce5c165295fbd895060-Paper-Conference.pdf

123. James Zou | Papers With Code, accessed May 2, 2025, https://paperswithcode.com/search?q=author%3AJames+Zou&order_by=stars

124. A Survey on Hallucination in Large Language and Foundation Models - Preprints.org, accessed May 2, 2025, https://www.preprints.org/manuscript/202504.1236/v1/download?ref=promptengineering.org

125. Mastering Generative AI Models: Trust & Transparency - Lumenova AI, accessed May 2, 2025, https://www.lumenova.ai/blog/generative-ai-models-ai-trust-ai-transparency/

126. AI Explained: Metrics to Detect Hallucinations - YouTube, accessed May 2, 2025, https://www.youtube.com/watch?v=a0jTPXwnRKs

127. Measuring GenAI adoption in the enterprise: Key performance metrics - Outshift | Cisco, accessed May 2, 2025, https://outshift.cisco.com/blog/genai-performance-metrics-measure-adoption

128. 4 Ways AI Content Detectors Work To Spot AI - Surfer SEO, accessed May 2, 2025, https://surferseo.com/blog/how-do-ai-content-detectors-work/

129. AI Drift | Microsoft Learn, accessed May 2, 2025, https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/mlops/drift-overview

130. What Is AI Alignment? - IBM, accessed May 2, 2025, https://www.ibm.com/think/topics/ai-alignment

131. Measuring Human-AI Value Alignment in Large Language Models, accessed May 2, 2025, https://ojs.aaai.org/index.php/AIES/article/download/31703/33870/35767

132. VisAlign: Dataset for Measuring the Alignment between AI and Humans in Visual Perception, accessed May 2, 2025, https://neurips.cc/virtual/2023/poster/73563

133. Getting AIs working toward human goals: Study shows how to measure misalignment, accessed May 2, 2025, https://cio.economictimes.indiatimes.com/news/artificial-intelligence/getting-ais-working-toward-human-goals-study-shows-how-to-measure-misalignment/120329734

134. Getting AIs working toward human goals: Study shows how to measure misalignment, accessed May 2, 2025,

https://m.economictimes.com/tech/artificial-intelligence/getting-ais-working-toward-human-goals-study-shows-how-to-measure-misalignment/articleshow/120301521.cms

135. AI alignment - Wikipedia, accessed May 2, 2025, https://en.wikipedia.org/wiki/AI_alignment

136. Measuring AI Model Performance: Tokens per Second, Model Sizes, and Inferencing Tools, accessed May 2, 2025, https://openmetal.io/resources/blog/ai-model-performance-tokens-per-second/

137. What are Tokens? - the docs - Voiceflow, accessed May 2, 2025, https://docs.voiceflow.com/docs/what-are-tokens

138. Tokens and Tokenization: Understanding Cost, Speed, and Limits with OpenAI's APIs, accessed May 2, 2025, https://www.prompthub.us/blog/tokens-and-tokenization-understanding-cost-speed-and-limits-with-openais-apis

139. Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy - PubMed Central, accessed May 2, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11852728/

140. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking, accessed May 2, 2025, https://www.mdpi.com/2075-4698/15/1/6

141. Towards Decoding Developer Cognition in the Age of AI Assistants - arXiv, accessed May 2, 2025, https://arxiv.org/html/2501.02684v1

142. The Impact of Implementing AI-Generated Audio Transcriptions on English Majors' Cognitive Load - ResearchGate, accessed May 2, 2025, https://www.researchgate.net/publication/388444730_The_Impact_of_Implementing_AI-Generated_Audio_Transcriptions_on_English_Majors'_Cognitive_Load

143. Full article: Evaluation of AI-generated reading comprehension materials for Arabic language teaching - Taylor & Francis Online, accessed May 2, 2025, https://www.tandfonline.com/doi/full/10.1080/09588221.2025.2474037?src=

144. Bias recognition and mitigation strategies in artificial intelligence healthcare applications - PMC - PubMed Central, accessed May 2, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11897215/

145. Managing the Two Faces of Generative AI | MIT CISR, accessed May 2, 2025, https://cisr.mit.edu/publication/2024_0901_GenAI_VanderMeulenWixom

146. Generative AI in Academic Research: Perspectives and Cultural Norms, accessed May 2, 2025, https://research-and-innovation.cornell.edu/generative-ai-in-academic-research/

147. Responsible governance of generative AI: conceptualizing GenAI as complex adaptive systems | Policy and Society | Oxford Academic, accessed May 2, 2025, https://academic.oup.com/policyandsociety/article/44/1/38/7965776

148. Declaration of Generative AI in Scientific Writing - Society for Vascular Surgery, accessed May 2, 2025, https://vascular.org/vascular-specialists/research/journals/declaration-generative-ai-scientific-writing

149. Generative AI Can Supercharge Your Academic Research - Harvard Business Publishing, accessed May 2, 2025, https://hbsp.harvard.edu/inspiring-minds/generative-ai-can-supercharge-your-academic-research

150. Controlling Large Language Model Outputs: A Primer | Center for Security and Emerging Technology, accessed May 2, 2025, https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/

151. AI We Can Trust: Controlling Generative AI to Ensure Reliable Creation - YouTube, accessed May 2, 2025, https://www.youtube.com/watch?v=uY7xIa77N8k

152. [2405.09794] Human-AI Safety: A Descendant of Generative AI and Control Systems Safety, accessed May 2, 2025, https://arxiv.org/abs/2405.09794

153. AAAI-24 Tutorial and Lab List, accessed May 2, 2025, https://aaai.org/aaai-24-conference/aaai-24-tutorial-and-lab-list/

154. Qiang Yang | Papers With Code, accessed May 2, 2025, https://paperswithcode.com/search?q=author%3AQiang+Yang&order_by=stars

155. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks, accessed May 2, 2025, https://www.jmlr.org/papers/volume22/21-0366/21-0366.pdf

156. PaLM 2 Technical Report - Google AI, accessed May 2, 2025, https://ai.google/static/documents/palm2techreport.pdf

157. DAN ROTH - Penn CIS - University of Pennsylvania, accessed May 2, 2025, https://www.cis.upenn.edu/~danroth/Research/cv.pdf

158. Factuality of Large Language Models: A Survey - arXiv, accessed May 2, 2025, https://arxiv.org/html/2402.02420v3

159. Factuality of Large Language Models: A Survey - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.emnlp-main.1088.pdf

160. [2402.02420] Factuality of Large Language Models: A Survey - arXiv, accessed May 2, 2025, https://arxiv.org/abs/2402.02420

161. Factuality of Large Language Models: A Survey - ACL Anthology, accessed May 2, 2025, https://aclanthology.org/2024.emnlp-main.1088/

162. Long-form factuality in large language models - Google DeepMind, accessed May 2, 2025, https://deepmind.google/research/publications/85420/