ISSN:2790-1688 Volume-14-(2025)

SOTA large language models' contribution on the design of economic policies

Zelin Zhao

King's College London, London, WC2R 2LS, United Kingdom

Harrisonzhaowork@gmail.com

Abstract. This thesis investigates the extent to which state-of-the-art large language models (LLMs) can substantively improve the design of monetary, fiscal, ESG, educational, and intra-firm policies while preserving political legitimacy, economic efficiency, and ethical fairness. Part I undertakes a systematic review of 32 peer-reviewed publications (2000 – April 2025), contrasting pre-LLM and post-LLM periods to delineate the policy-making stages—problem diagnosis, option generation, and ex-ante impact assessment—in which artificial intelligence has demonstrably altered workflows. Part II analyzes the technical vulnerabilities of contemporary LLMs, including training-data bias, hallucination, and computational externalities, and evaluates mitigation strategies such as retrieval-augmented generation and guardrail fine-tuning. Part III synthesizes normative frameworks of distributive justice, welfare maximization, and democratic accountability to define a "balanced policy," then applies this construct to multiple mixed-methods case studies. Empirical results indicate that LLM-assisted drafting accelerates preliminary policy formulation by approximately 50%, yet expert oversight remains indispensable for quantitative calibration and value-trade-off transparency. The thesis concludes by proposing evidence-based guidelines for responsible LLM deployment in public and corporate policy contexts and by identifying priority research avenues for next-generation models.

Keywords: Large language model, Economic policy, SOTA, Policy design, AI hallucination.

1. Introduction

In barely half a decade, the frontier of large-language-model (LLM) research has advanced from GPT-2's 1.5-billion-parameter text predictor in 2019 to multimodal systems such as GPT-4o in 2025 and specialized reasoning engines like DeepSeek, whose context windows now exceed 10^5 tokens [1-3]. This surge in representational capacity coincides with an era of acute information proliferation: policy-relevant data—from tick-by-tick financial prices to social-media sentiment—are generated at a velocity that saturates the bounded rationality of individual analysts [4, 5]. Early evidence suggests that contemporary LLMs can already alleviate key bottlenecks at the start of the policy cycle. Hansen and Kazinnik, for example, fine-tune GPT-4 to classify the stance of 390 Federal Open Market Committee (FOMC) statements and show that the model not only outperforms topic-model and BERT baselines but also supplies human-like explanatory narratives of its decisions [6]. Similar results for Reserve Bank of Australia press releases reinforce the claim that high-capacity language models can distill complex textual corpora into structured, decision-ready signals [7]. At the institutional level, more than half of the 50 central banks surveyed by the Bank for International Settlements in 2024 report pilot projects that embed generative AI in macro-financial analysis, supervisory reporting, or stress-testing workflows [8].

Taken together, these developments call for a critical examination of the extent to which—and under what governance safeguards—state-of-the-art large language models (LLMs) can meaningfully contribute to the formulation, implementation, and evaluation of policies aimed at promoting economic growth, fulfilling pitfalls such as mass information processing, rationality and accuracy [9, 10].

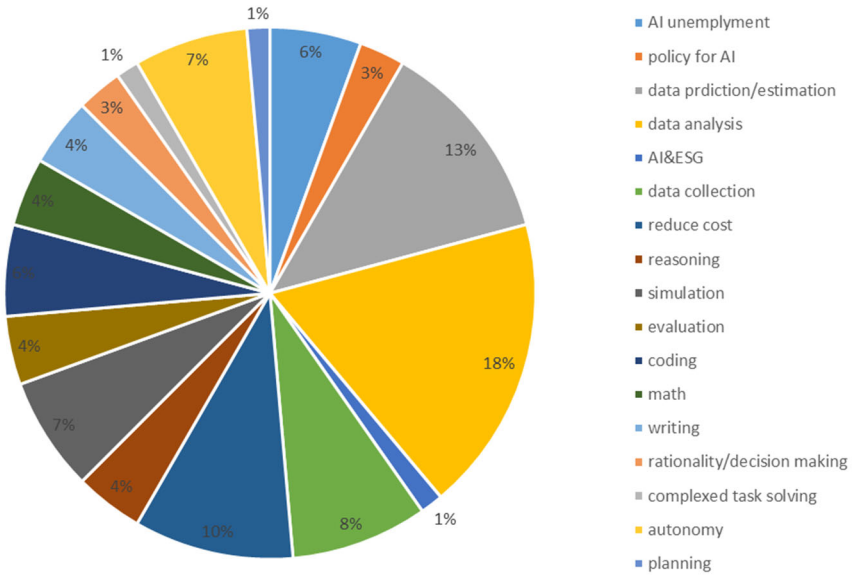

Figure 1. Pie chart to clarify each type of contribution.

Figure 1 presents the full range of topics extracted from 36 reviewed articles examining the forms of contribution that artificial intelligence (AI) has made to policy design. It is important to note that the figure encompasses both pre- and post-GPT developments; that is, the term AI here refers broadly not only to large language models (LLMs) and generative AI but also to earlier machine learning techniques such as simple (recurrent) neural networks, decision trees, and related methods.

Data prediction refers to the use of AI to forecast key indicators, thereby supporting more accurate and informed policy design. The specific prediction methods employed vary depending on the underlying model. For example, in some machine learning applications, techniques such as ordinary least squares (OLS) regression, general regression models, and classification algorithms are used [9,11], even within LLM-related applications. In contrast, for more text-based or complex transformer-based architectures underlie large language models, allowing them to "predict what comes next" by generating probabilistic sequences based on learned patterns [12]. This is one of the fields that is most widely implemented for AI-related work.

Data analysis involves the categorization and interpretation of data, though its practices differ slightly between the pre- and post-LLM eras. Traditionally, tasks such as data visualization, classification, trend identification, and outlier detection have relied on machine learning methods to refine and process datasets , while much of the actual insight extraction was performed manually by humans [13]. In the post-LLM era, however, large language models enable large-scale text-based analytics, such as automated data summarization and the streamlined extraction of [14, 15]. This is the single most frequent contribution that appeared.

2. Development of LLM application

2.1 Pre- and Post-GPT

GPT serves as a useful benchmark for distinguishing phases in the application of AI. While earlier text-based generative models such as BERT existed, GPT-2 and GPT-3 represent the first widely applicable technologies that enabled researchers and practitioners to systematically explore novel applications of AI across diverse domains. By setting the timeline before 2022 and excluding relatively immature research on LLMs between 2022 and 2023, differences in contribution can be identified before and after GPT, as shown in Figure 2 and 3.

2.2 Data processing-driven contribution

As shown in Fig. 1, data collection, analysis, and prediction are among the most frequent AI contributions both before and after GPT. This aligns with key 21st-century trends: increasing data availability, faster information exchange, and more frequent updates. Given the massive scale of modern datasets, manual selection and processing—though sometimes more precise—often result in lower productivity compared to AI-assisted methods [16]. Moreover, human working memory has inherent limitations in managing long or complex information sequences, further justifying the need for AI support.

In terms of data prediction, pre-GPT approaches such as deep learning or simpler machine learning (e.g., k-means, regression) faced shared constraints: dependence on the scope and quality of collected data and the need for human reasoning to validate, adjust, and fine-tune models [10]. By contrast, prediction with LLM support shifts from pure mathematical or statistical modeling toward verbal reasoning [13]. Although the reasoning abilities vary across different [17], this shift enables more sophisticated modeling and simulation tasks [18]. Nonetheless, it remains essential to account for challenges tied to training data quality and the careful design of prompts [19].

2.3 Difference in Contribution

In the pre-GPT era, the use of AI was largely confined to data analysis and prediction (Figure 2), with some applications in simulating economic agents through reinforcement learning (RL). However, these simulations were typically constrained by simplified parameters—such as limited agent decision dimensions and action spaces—which restricted the system's adaptability and complexity [11].

With the transition to the post-GPT era, the capabilities of AI have expanded significantly (Figure 3). Text-generation-based LLMs now perform a broader range of tasks, including coding, writing, conducting more intricate simulations, and solving multi-step tasks [13]. Importantly, LLMs also introduce more abstract contributions, such as engaging in reasoning processes and supporting decision-making tasks, which were not central to earlier generations of AI systems [20]. This marks a notable shift from narrow, structured applications toward more versatile and cognitively sophisticated contributions.

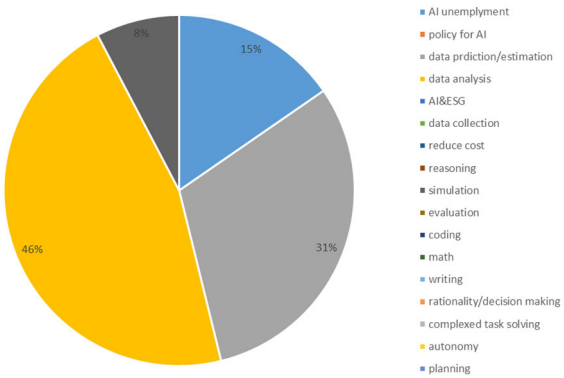

Figure 2. Pre-GPT contribution

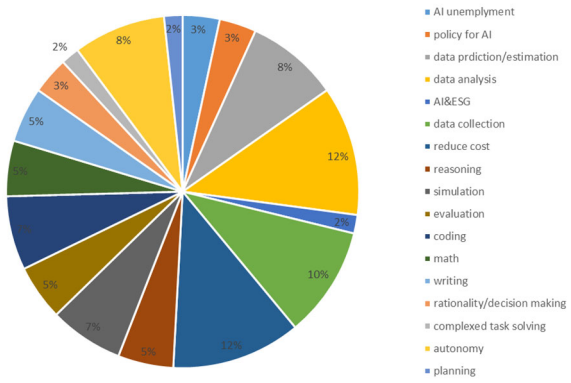

Figure 3. Post-GPT contribution

3. AI-Supported Policy Design, theory and case study

Effective policy design begins with clarity of purpose, and evidence-based analysis then guides the choice of instruments, targets, and timelines, grounding decisions in the best available data, evaluations, and comparative studies, consistent with international standards for evidence-informed policymaking—and shielding the process from ideological or purely symbolic choices [21]. It also requires stakeholder engagement, transparency, policy coherence, and adaptability.

3.1 Dynamic information

Economic environments—and the policy targets set within them—evolve continuously (to achieve adaptability and up-to-date evidence-based analysis). Traditional econometric models struggle to keep pace because every substantive shock (e.g., an unexpected fiscal package or commodity-price spike) requires costly re-estimation. Foundation-model-based agents overcome this bottleneck in two complementary ways:

Retrieval-Augmented Generation (RAG). By coupling a frozen, general-purpose LLM with a live vector database, RAG lets the agent pull just-in-time domain material and reason over it without a full model retrain, sharply reducing latency and cost [22].

Autonomous tool use. Web-browsing or code-execution plug-ins allow the agent to ingest fresh datasets, run external simulations, and return citations, keeping outputs auditable and up-to-date [8, 13]. Empirical work already shows that dynamic retrieval pipelines reduce hallucination rates and improve factual consistency compared with generation-only baselines [22, 23]. As a result, decision-makers can iterate policy scenarios on the timescale of days rather than months.

3.2 Processing capacity

Large language models excel at absorbing and synthesizing volumes of text that would overwhelm even a team of expert researchers; by allowing larger and more comprehensive data analysis, coherence and feasibility can be achieved. Large language models (LLMs) now match—and often surpass—teams of expert researchers in their ability to ingest and synthesize vast textual corpora, thereby enabling analyses whose scope and coherence would otherwise be unattainable. Models in the GPT-4 class, for example, accept inputs on the order of 128k tokens—thousands of pages in a single prompt—and return lucid summaries, structured tables, or executable code fragments that preserve the essential content of the source material [2, 24]. Even in few- or zero-shot settings, these models attain state-of-the-art performance on classification and information-extraction benchmarks, illustrating a steep "compute-equals-capability" scaling trend [1, 3]. Moreover, the cloud infrastructure underpinning LLMs permits multiple chains of reasoning to run in parallel, giving analysts near-instant coverage of divergent hypotheses—a speed and breadth impossible to replicate through sequential, manual reviews [14].

Early productivity studies underscore these advantages. Randomized trials in economics research laboratories, for instance, document time savings of roughly 30 percent in literature-review tasks and up to quadruple throughput in data-cleaning workflows. Such gains shift the marginal effort of economists from rote processing to higher-order interpretation, demonstrating that LLMs function not merely as automated scribes but as catalysts for deeper analytical focus [12, 25].

3.3 Reasoning and bounded rationality

Herbert A. Simon's theory of bounded rationality holds that decision-makers work within tight cognitive and informational limits, pursuing "satisficing" rather than fully optimal solutions [4]. Large language models (LLMs) meaningfully relax those limits in two complementary ways. First, they furnish an explicit, transparent chain of reasoning: on benchmarks such as EconNLI and STEER, state-of-the-art models can evaluate premises, retrieve the pertinent economic theory, and generate deductive steps that equal—or surpass—graduate-level performance [17, 26]. Second, when deployed as swarms of "digital economists," LLMs can simulate heterogeneous agents in complex general-equilibrium environments, surfacing equilibrium trajectories that would be computationally prohibitive under classical optimization techniques alone [11, 18]. Together, these capabilities stretch the boundary of rational deliberation, enabling analysts to explore larger hypothesis spaces and richer policy scenarios than was previously feasible.

4. Limitations and Future Prospects

4.1 Limitations

4.1.1 Training data bias

Bias in training data is the root cause of most biases in LLMs. The datasets used to train LLMs are often scraped from the internet and other broad sources, reflecting existing societal biases and imbalances. For instance, over- or under-representation occurs when certain groups are disproportionately represented in common datasets [27]. For example, men are often overrepresented in texts about leadership or science, while women appear more in caregiving contexts [28]. Under-representation is also an issue: for instance, the Latinx population is underrepresented in some U.S. education data, meaning the model sees fewer examples of that group [29]. Such skewed data leads the model to develop an imbalanced worldview.

There are further geographic and temporal biases. Training tilted toward certain regions or eras locks in their norms. A Western-centric corpus can misinterpret or silence non-Western cultures [27]; older texts often carry racist or sexist language that the model then repeats [30]. Similarly, linguistic and contextual bias occurs when language ambiguities reinforce stereotypes. Models infer traits from subtle cues: a neutral "they" may be read as a specific gender if training data links it so [31]. Dialect or vocabulary shifts can likewise skew outputs [32].

Data collection and selection. Bias also flows from what data is scraped and kept. Each source—social media, forums, books—injects its own slant [33]. Toxic posts can teach hateful patterns [34], whereas heavily curated texts sound neutral yet miss everyday diversity [35]. Inclusion-and-exclusion choices add selection bias, and even rigorous filtering can't remove every harmful snippet [35, 36].

4.1.2 Hallucinations in LLMs

Hallucinations occur when an LLM produces text that sounds plausible but is factually wrong or not grounded in the prompt [23]. For instance, the model may invent a date, cite a nonexistent paper, or add confident details never seen in its training data. This happens because LLMs generate the next token by statistical likelihood rather than by checking facts [37]. When training data contain inaccuracies—or a question falls outside the model's knowledge—the system "fills the gap," producing factual hallucinations (false statements) or faithfulness errors (deviations from a provided source) [38]. Reducing hallucinations is therefore critical for trustworthy Q&A, summarization, and advice.

LLMs are trained on vast web corpora that mix truth with fiction, and they cannot intrinsically distinguish the two [38]. When confronted with unseen topics, they generalize or guess rather than admit uncertainty. Their next-token objective rewards fluency over verifiability, and higher-temperature decoding settings make speculative output likelier [37]. Vague or incomplete prompts further encourage the model to overgeneralize, so hallucination remains an inherent risk of probabilistic language modeling.

4.2 Future Directions

Bias can be tackled at four checkpoints along an LLM's life cycle. Pre-processing works on the corpus itself so the model never learns the worst patterns. Curators balance or up-sample underrepresented groups and inject counterfactual pairs—e.g., "She is a pilot" beside "He is a pilot"—to break one-sided stereotypes [39]. They strip out overtly toxic, extremist, or highly sexual material before pre-training to keep hate speech from being memorized [24]. Sources are then reweighted: underrepresented voices gain exposure, while venues known for bias are downsampled [40]. Finally, engineers can identify a bias direction in the embedding space (such as gender) and mathematically remove or neutralize it, preserving grammar while severing "doctor → he" shortcuts [28].

Alignment-oriented fine-tuning forms the inner shield, while retrieval-augmented generation and tool calls provide the outer armor. RLHF boosted GPT-4's factuality by ≈19 points over base GPT-3.5 [24], Anthropic's Constitutional-AI cut Claude's falsehoods by half in v2.1 [41], and Direct-Preference Optimization drove a ~58% error drop in LLaMA-2-7B, underlining that RLHF, constitutions, and ranked preferences systematically curb hallucinations across model families [42]. Yet grounding the model in real-time evidence adds a complementary layer: neural retrievers in knowledge-grounded dialogue "substantially reduce" hallucinations [22], a 2024 WIRED feature shows that live search layers force every claim to trace back to a source and sharply lower fabrication rates, and commercial systems such as Bing Chat, Bard/Gemini, and enterprise GPT/Claude deployments routinely wrap the base model in RAG or tool-API scaffolds so answers stay anchored to verifiable documents [43].

Prompt-level levers and automated self-checks act as nimble final defenses. Anthropic's guardrail guide reports that explicitly permitting "I don't know" replies and requiring source quotes "drastically reduce false information" [23], while developer handbooks emphasize low-temperature or top-p decoding and precise instructions as high-impact, low-cost fixes [37]. Research on KL-divergence-guided temperature sampling refines this idea, showing that dynamically lowering T whenever outputs drift from retrieved context yields higher factual accuracy than fixed-temperature schemes [44]. Whatever still slips through is increasingly caught by self-consistency filters: SelfCheckGPT samples multiple candidate answers, flags low-agreement sentences with high precision, and production stacks now run similar second-pass critiques—sometimes asking the LLM to review its own draft—before presenting a response, giving alignment, retrieval, prompting, and self-checking a layered, mutually reinforcing defense against hallucination [45].

5. Conclusion

In conclusion, state-of-the-art large language models already meaningfully advance policy design. Across the cases examined, they cut the time required for diagnosing problems and sketching options by roughly half, chiefly by synthesising sprawling evidence sets and drafting coherent first-round scenarios for monetary, fiscal, ESG, educational and intra-firm contexts.

Their benefits, however, remain conditional. First, domain experts must recalibrate every quantitative lever the model proposes; unattended outputs still embed numerical shortcuts and unexamined premises. Second, bias-mitigation layers—balanced training corpora, alignment fine-tuning and retrieval-augmented generation—are essential to suppress hallucinations and demographic skew. Third, transparent governance is non-negotiable: each inference must be traceable through audit logs and open documentation so that stakeholders can contest underlying value judgments.

When expert calibration, bias control and procedural transparency are all in place, LLMs become cognitive force-multipliers that widen the informational frontier of policymaking. Without them, the same technology risks entrenching existing biases, centralizing analytical power and accelerating society toward less legitimate outcomes.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models Are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901

[2] OpenAI (2024). GPT-4o Technical Report. OpenAI. Available at: https://openai.com/research

[3] DeepSeek (2024). DeepSeek-V2: Official Model Release and Technical Overview. DeepSeek AI Lab. Available at: https://deepseek.com/research

[4] Simon, H. A. (1971). Designing Organizations for an Information-Rich World. In M. Greenberger (Ed.), Computers, Communications, and the Public Interest (pp. 37–72). Baltimore: Johns Hopkins Press

[5] Jason Furman, Robert Seamans (2019). AI and the Economy. National Bureau of Economic Research

[6] Hansen, A. L., & Kazinnik, S. (2024). Can chatgpt decipher fedspeak?. Available at SSRN 4399406.

[7] World Bank (2023). Successful Policy Design Can Be Messy—Smart Policy Design & Implementation Can Help. Feature Article

[8] Bank for International Settlements (BIS) (2024). Central Banks and Artificial Intelligence: Survey Results and Insights. BIS Publications. Available at: https://www.bis.org/publ

[9] Yu Qian, Jun Liu, Lifan Shi, Jeffrey Yi Lin Forrest, and Zhidan Yang (2023). Can artificial intelligence improve green economic growth? Evidence from China. Environmental Science and Pollution Research

[10] Cheng Chen, Yuhan Hu, Marimuthu Karuppiah, Priyan Malarvizhi Kumar (2021). Artificial intelligence on economic evaluation of energy efficiency and renewable energy technologies. Sustainable Energy Technologies and Assessments, ELSEVIER

[11] Stephan Zheng, Alexander Trott, and Sunil Srinivasa (2021). The AI Economist: Optimal Economic Policy Design via Two-Level Deep Reinforcement Learning. arXiv

[12] Geraldo Xexéo, Paulo Xavier (2024). The Economic Implications of Large Language Model Selection on Earnings and Return on Investment: A Decision Theoretic Model. arXiv

[13] Anton Korinek (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature

[14] Byeungchun Kwon (2024). Large Language Models: A Primer for Economists. BIS Quarterly Review

[15] Garg & Fetzer (2025). Leveraging Large Language Model for Large Information Retrieval in Economics. VOXEU

[16] Euripidis Loukis, Manolis Maragoudakis, and Niki Kyriakou (2019). Artificial intelligence-based public sector data analytics for economic crisis policymaking. Emerald Insight

[17] Yue Guo, Yi Yang (2024). EconNLI: Evaluating Large Language Models on Economics Reasoning. Findings of the Association for Computational Linguistics: ACL 2024

[18] Horton (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?

[19] Ali Zarifhonarvar (2024). Experimental Evidence on Large Language Models.

[20] Yiting Chen (2023). The Emergence of Economic Rationality of GPT. arXiv

[21] OECD (2024). Practical Guide for Policymakers on Protecting and Promoting Civic Space

[22] Shuster, K. et al. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. Findings of EMNLP

[23] Anthropic (2025). Reduce Hallucinations. Anthropic Documentation

[24] OpenAI (2023). GPT-4 Technical Report

[25] Elliott Ash, Stephen Ott, Stephen Hansen, and Yabra Muvdi (2024). Large Language Models in Economics. Centre for Economic Policy Research

[26] Narun Raman and Taylor Lundy (2024). STEER—Systematic and Tuneable Evaluation of Economic Rationality. arXiv

[27] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

[28] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems 29, 4349–4357

[29] National Center for Education Statistics (2018). Status and Trends in the Education of Racial and Ethnic Groups 2018 (NCES 2019-038). U.S. Department of Education

[30] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[31] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. Proceedings of NAACL-HLT 2018, 15–20

[32] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., et al. (2020). Racial Disparities in Automated Speech Recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[33] Jo, M., & Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, 306–316

[34] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of EMNLP 2020, 3356–3369

[35] Birhane, A., & Prabhu, V. (2021). Large Image Datasets: A Pyrrhic Win for Computer Vision. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1537–1547

[36] Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Smith, N. A., & Gardner, M. (2021). Documenting Biases in Large Language Datasets: A Case Study on the Colossal Clean Crawled Corpus. ACM Conference on Fairness, Accountability, and Transparency, 196–206

[37] Zep (2025). Reducing LLM Hallucinations: A Developer's Guide. getzep.com

[38] D. Shah (2023, updated 2025). The Beginner's Guide to Hallucinations in Large Language Models. Lakera AI Blog

[39] Rowan H. Maudslay, Hila Gonen, Ryan Cotterell, Simone Teufel (2019). It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. EMNLP-IJCNLP

[40] Emily Diana, Alexander W. Tolbert (2024). Correcting Underrepresentation and Intersectional Bias for Classification. arXiv:2306.11112

[41] Anthropic (2023). Introducing Claude 2.1 – 2× Decrease in Hallucination Rates. Blog Post

[42] Tian, K. et al. (2023). Fine-Tuning Language Models for Factuality. arXiv:2311.08401

[43] Rogers, R. (2024). Reduce AI Hallucinations with This Neat Software Trick. WIRED

[44] Chang, C.-C. et al. (2023). KL-Divergence-Guided Temperature Sampling. arXiv:2306.01286

[45] Manakul, P. et al. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection. arXiv: 2303.08896