# Prompt Position Optimization for Improving Hallucination Detection and Mitigation in Generative AI

Yash Ganguly
School of Computing, Clemson University, Clemson, SC, USA.
yashganguly74@clemson.edu

Vearun Keahli
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
kohlivarun@binghamton.edu

Greant Wealters
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
grant.walters262@unr.edu

**Abstract**

Generative artificial intelligence systems, particularly large language models, have demonstrated remarkable fluency across diverse domains, yet their propensity to produce factually incorrect or contextually unsupported content—commonly termed hallucination—continues to undermine trustworthiness in high-stakes applications. While substantial research has focused on detecting and mitigating hallucinations through post-hoc verification, retrieval augmentation, and fine-tuning strategies, the structural role of prompt positioning within input contexts remains underexplored as a systemic lever for enhancing detection and mitigation. This paper presents a comprehensive systems-oriented analysis of prompt position optimization as a cross-cutting mechanism for improving hallucination safeguards in generative AI. We examine how positional biases, including recency and lost-in-the-middle effects, shape model attention and information fidelity, and argue that dynamic, context-aware placement of verification, calibration, and steering prompts can substantially improve hallucination detection rates while reducing computational overhead. Drawing on architectural, infrastructural, and governance perspectives, the paper situates prompt position optimization within a broader socio-technical framework that encompasses fairness, robustness, deployment scalability, and sustainability. We discuss architectural design patterns that integrate learnable prompt insertion policies alongside continual monitoring pipelines, and evaluate trade-offs between latency, cost, and detection performance. Further, we address governance implications, including transparency in prompt interventions and the need for standardized audit protocols that account for prompt position as a controllable design parameter. By synthesizing insights from prompt engineering, model alignment, and systems engineering, this work establishes prompt position optimization as a foundational component of resilient, accountable generative AI infrastructure and charts pathways for interdisciplinary research and regulatory engagement.

**Keywords**

prompt position optimization, hallucination detection, generative AI, large language models, system architecture, governance, robustness, fairness, infrastructure, sustainability.

## 1. Introduction

The rapid integration of generative artificial intelligence into healthcare, legal analysis, scientific research, and public administration has elevated the importance of content veracity to a matter of institutional and societal concern. Hallucination, defined as the generation of plausible-sounding yet factually incorrect or ungrounded outputs, remains a persistent failure mode even in state-of-the-art large language models (LLMs) [1], [2]. Despite advances in pre-training on expansive corpora, reinforcement learning from human feedback, and retrieval-augmented generation, models continue to exhibit context-unfaithful behavior, especially under long input conditions, ambiguous queries, or adversarial prompts [3]. As regulatory frameworks such as the European Union's AI Act and the U.S. Executive Order on Safe, Secure, and Trustworthy AI increasingly call for verifiable safety mechanisms, detecting and mitigating hallucination has become an urgent systems-level challenge that intersects technical design, operational governance, and public accountability.

Existing approaches to hallucination management broadly divide into detection pipelines—such as self-consistency checks, uncertainty quantification, and external fact verification—and mitigation techniques that modify model behavior through controlled decoding, knowledge grounding, or prompt engineering [4], [5]. While these methods have made measurable progress, they typically treat the prompt as a static, monolithic component whose textual content and semantic framing are optimized, but whose spatial positioning within the input sequence receives minimal systematic attention. Yet a growing body of evidence demonstrates that language models exhibit strong positional biases: information presented near the beginning or end of a prompt is processed with higher fidelity, while critical content placed in the middle is frequently overlooked, a phenomenon known as the lost-in-the-middle effect [3]. These biases directly influence whether oversight instructions, factual anchors, or verification cues are effectively utilized by the model, creating a substantial, and largely unexamined, attack surface for hallucination.

This paper advances the thesis that prompt position optimization—the principled, dynamic selection of where to insert steering, calibration, and verification prompts within an input context—constitutes a powerful, systemically coherent strategy for improving hallucination detection and mitigation. Unlike mere prompt rewriting, position optimization operates at the architectural layer between the application interface and the underlying generative model, offering a non-invasive mechanism that can be overlaid onto existing models and workflows. By repositioning hallucination-detection prompts to exploitation points where attention weights are strongest, and by adaptively inserting mitigation prompts based on run-time uncertainty signals, systems can achieve substantial gains in detection sensitivity and reduction in factual errors without the prohibitive costs of full model retraining. The concept draws on insights from prompt tuning, in-context learning, and selective prompt insertion, yet extends these paradigms toward the specific goals of reliability and safety at the scale of deployed infrastructure.

In the sections that follow, we examine the limitations of current hallucination detection and mitigation strategies through a positional lens, formalize the design space of prompt position optimization, and propose architectural frameworks that integrate position-aware prompt management into continuous monitoring and response loops. We then analyze system-level trade-offs in latency, throughput, fairness, and sustainability, and outline governance

considerations necessary for accountable deployment. Through this analysis, we seek to establish prompt position optimization as a foundational component of trustworthy generative AI systems and to inform both engineering practice and policy development.

## 2. Background and Related Work

Hallucination in generative AI has been comprehensively surveyed, with taxonomies differentiating between intrinsic hallucinations that contradict source material and extrinsic hallucinations that cannot be verified against any provided context [1]. Detection methods span a broad spectrum from black-box approaches such as SelfCheckGPT, which uses multiple stochastic samples to identify inconsistencies, to white-box techniques that scrutinize internal model states such as attention weights and hidden representations [2], [6]. Simultaneously, mitigation strategies have incorporated knowledge grounding via retrieval-augmented generation, where external documents are appended to the prompt, and various decoding-time interventions such as constrained beam search or factual nucleus sampling [5]. While these methods operate on the content of prompts, they do not explicitly optimize where within the tokenized sequence that content resides.

Prompt engineering has matured into a recognized discipline, with structured frameworks cataloging techniques such as chain-of-thought reasoning, few-shot exemplar selection, and persona assignment [7]. Work on chain-of-thought prompting has demonstrated that explicitly instructing models to articulate intermediate reasoning steps before concluding substantially improves veracity on reasoning-intensive tasks, positioning such reasoning segments at the end of a prompt to exploit recency bias implicitly [4]. However, systematic evaluation of prompt component placement across varying context lengths is recent and motivated by findings that LLM performance degrades severely for information located in the middle of long inputs, even when that information is critical to accurate response generation [3]. This recency and primacy bias suggests that factual anchors, verification rules, or instructions to cross-check generated claims should be deliberately placed to align with the model's attention profile.

Selective insertion of prompts into input sequences has been explored in the context of prompt tuning for task adaptation. One line of work learns to insert soft prompt tokens at positions chosen by a lightweight policy network, demonstrating that adaptive placement outperforms fixed prefix or suffix insertion on multiple natural language understanding benchmarks [9]. Although developed for general task performance, this learning-to-insert paradigm provides a blueprint for hallucination-oriented prompt placement, where the objective shifts from task accuracy to factual consistency and detection recall. Similarly, adversarial robustness research has revealed that subtle reordering of prompt components can drastically change model susceptibility to injection attacks and factual drift, underscoring the security dimension of prompt position [8]. These developments collectively motivate a system-level treatment of prompt position optimization as a safety-critical control variable, rather than a peripheral design choice.

## 3. Positional Biases and Their Impact on Hallucination Detection

To understand the leverage that prompt positioning provides, it is essential to appreciate the interaction between transformer attention mechanisms, context length, and information fidelity. In autoregressive transformer models, each token's representation is aggregated from preceding tokens through self-attention, creating an inductive bias where tokens at the temporal extremes of a sequence often receive disproportionately high attention weights. This

results in a U-shaped recall curve across the context window: information presented at the very beginning and the very end is recalled more accurately than information in intermediate positions. The effect intensifies as context windows expand, with models demonstrating nearly chance-level retrieval accuracy for facts placed around the midpoint of sequences extending to tens of thousands of tokens [3]. For hallucination detection, this implies that a verification instruction appended as a prefix or suffix may function effectively, while an identical instruction embedded in the middle of a multi-document retrieval prompt or a lengthy conversation history is likely to be functionally invisible to the model.

The consequences for detection are direct and severe. Consider a deployment scenario in which a healthcare summarization system is required to verify generated statements against integrated patient records placed within a long context. If a "verify against the provided records and flag any unsupported claims" directive is positioned amid the clinical notes rather than at the end, the model may generate summaries without attending to the verification constraint, leading to undetected factual errors. Similarly, in legal document analysis, a calibration prompt designed to prompt the model to express uncertainty or request clarification can be rendered inert if it is inserted at a position that coincides with an attention trough. Empirical studies on prompt robustness confirm that model outputs can change dramatically under semantically equivalent reorderings, indicating that positional decisions function as a hidden hyperparameter with safeguarding consequences [8].

Beyond detection, positional bias also affects mitigation strategies that rely on in-context learning to shape model behavior. Few-shot examples of refusing to answer unverified queries, or demonstrations of citing retrieved evidence, must occupy positions that maximize their influence on subsequent generation. If such calibration examples are buried within a dense prompt, the model may default to its pre-trained tendencies toward overconfident synthesis. Thus, position optimization emerges as a critical lever for operationalizing model alignment and safety protocols within the input assembly pipeline, directly bridging low-level architectural properties with high-level trustworthiness objectives.

## 4. Prompt Position Optimization as a System Strategy

Prompt position optimization can be formalized as a dynamic scheduling problem: given an input context, a set of available detection and mitigation prompt modules, and a model with known positional biases, select insertion positions that maximize a composite metric of factual accuracy, calibrated confidence, and resource efficiency. This formulation moves beyond static prompt templates toward a run-time adaptive mechanism that treats position as a first-class design parameter, alongside content and format. The design space includes global repositioning of all prompt components, selective insertion of lightweight sentinel tokens that activate attention, and learning-based policies that anticipate where informative signals are most needed based on input characteristics.

From a systems standpoint, implementing position optimization requires a middleware layer interposed between the application and the inference engine. This layer monitors context length, estimates attention saturation points, and employs heuristics or learned models to rearrange prompt segments while preserving semantic coherence. The simplest heuristic exploits the known U-shaped recall curve by placing verification and mitigation prompts as a suffix, ensuring proximity to the output token positions. However, suffix placement is not universally optimal. In chain-of-thought scenarios, inserting a fact-checking interlude between reasoning steps may counteract error propagation more effectively than a single end-of-prompt directive [4]. In multi-turn conversational systems, periodically re-inserting a

calibration prompt at turn boundaries can prevent dialogue drift without adding excessive latency. More sophisticated approaches leverage lightweight encoder modules that predict token importance scores, allowing selective insertion of soft prompt tokens at positions where they can maximally influence downstream attention, akin to the learned prompt insertion strategies explored for task tuning [9]. Extending such methods to hallucination safety involves training insertion policies on annotated corpora with factual consistency labels, enabling the system to learn that verification prompts placed just before a claim generation phase yield higher detection sensitivity.

The optimization objective must balance detection improvement against computational overhead. Each additional prompt token increases input length quadratically in transformer self-attention complexity, and repositioning prompts may require token re-encoding, adding latency in real-time settings. A viable infrastructure therefore couples position optimization with caching strategies for common prefix-suffix structures and employs speculative position scoring to abort unfavorable placements early. In cloud-scale deployments, the position optimizer becomes part of the inference orchestration layer, interacting with load balancers and token budgeting modules to deliver safety guarantees within service-level agreements. This elevates prompt position from a static configuration file entry to a dynamic, observable, and governable system property.

## 5. Architectural Frameworks for Adaptive Prompt Placement

Integrating prompt position optimization into production generative AI systems demands an architectural framework that supports observability, continuous improvement, and safety verification. We envision a modular design composed of a Context Analyzer, a Position Policy Engine, a Prompt Assembly Manager, and a Hallucination Feedback Collector. The Context Analyzer profiles incoming input for length, structure, domain, and estimated complexity, while also interfacing with model-internal signals such as attention entropy where feasible. The Position Policy Engine applies rules or trains online reinforcement learning policies to select insertion positions for detection and mitigation prompts, drawing on historical performance data and the estimated positional bias profile of the serving model. The Prompt Assembly Manager then constructs the final input sequence, respecting token budget constraints and ensuring that inserted prompts do not fracture critical semantic units. Finally, the Hallucination Feedback Collector captures downstream verification results—whether from human annotators, automated fact-checking modules, or self-consistency checks—and funnels them back to refine the position policy.

In such an architecture, the choice between rule-based versus learned position policies introduces system trade-offs. Rule-based policies offer determinism, low computational overhead, and auditability, making them suitable for regulated domains where every safety intervention must be explainable. For example, a rule might specify that verification prompts must always occupy the final 10 percent of the token budget and be duplicated at the halfway point for contexts exceeding two thousand tokens. Learned policies, by contrast, can adapt to nuanced task distributions and model-specific attention quirks, potentially uncovering non-obvious placements that yield superior hallucination detection. However, they introduce non-determinism and require careful validation to avoid emergent unsafe placements. A hybrid design that uses rules as guardrails overlaid on a learned core combines the benefits of both, with the rule layer ensuring that catastrophic positioning failures are precluded while the learned component fine-tunes for efficacy.

Deploying adaptive prompt placement also intersects with retrieval-augmented generation pipelines. In RAG architectures, retrieved documents are typically placed in the context window before the user query. Position optimization can reorder these documents so that the most relevant pieces border the query or the verification instructions, rather than simply appearing in retrieval rank order. It can also insert sentinel prompts that instruct the model to compare generated claims against specific retrieved passages excerpted at strategic positions. This tight coupling between document positioning and prompt placement creates a symbiotic relationship in which retrieval relevance and hallucination mitigation reinforce each other, turning the prompt assembly layer into a dynamic orchestrator of knowledge grounding and safety.

## 6. Governance, Fairness, and Policy Implications

Prompt position optimization, while technically situated within the inference pipeline, carries significant implications for governance, fairness, and public policy. As regulatory instruments demand transparency regarding the mechanisms used to ensure AI safety, position optimization surfaces as an algorithmic intervention that is simultaneously powerful and opaque to external auditors. A system that dynamically rearranges prompts to suppress hallucinations may also, inadvertently or by design, alter distributional outputs in ways that affect protected groups, factual framing, or the balance of perspectives. If a verification prompt is consistently positioned to favor certain types of sources or enforce a specific ontological framing, the optimization process could institutionalize subtle biases that escape standard evaluation suites.

Fairness must be assessed not only through aggregate hallucination rates but across demographic and topical slices. For instance, in question-answering systems serving diverse linguistic communities, prompt position policies trained primarily on English corpora may embed positional preferences that are suboptimal or harmful for low-resource languages, where syntax and discourse structure differ and attention biases may manifest differently. Continuous monitoring and disaggregated evaluation are therefore essential. An equitable deployment requires that position optimization modules are tested for differential efficacy and that feedback loops encompass representative annotator populations so that the learned insertion policy does not overfit to majority-group patterns of factual verification.

From a policy standpoint, prompt position optimization should be incorporated into the scope of AI system audits and impact assessments. Auditors would need access to logs documenting which prompts were placed where, at what time, and with what detected effect on output veracity, subject to privacy constraints. Standardized documentation artifacts akin to model cards or system cards could include a “position policy card” specifying the placement logic, the positional bias model assumed, and the validation methodology. Such transparency facilitates accountability while preserving intellectual property, as the policy engine’s architecture and hyperparameters can be disclosed without revealing proprietary model weights. Moreover, in high-risk sectors such as medical diagnosis or legal reasoning, regulation could mandate that position optimization be employed as part of a layered safety strategy, with defined performance thresholds for hallucination detection under standardized positional stress tests. This policy integration underscores that prompt position is not merely an engineering convenience but a governance instrument with direct influence on AI trustworthiness.

## 7. Deployment and Infrastructure Considerations

Bringing prompt position optimization to life in large-scale, multi-tenant inference environments introduces infrastructure challenges around latency, state management, and cost. Generative AI services routinely handle millions of requests per day with stringent tail-latency requirements. Introducing a position optimization middleware that incurs additional pre-processing, possibly involving auxiliary model inference for position scoring, must meet microsecond or low-millisecond budgets to remain practical. This demands highly optimized implementations that leverage model distillation for scoring policies, batch processing of concurrent requests, and tight integration with inference servers such as vLLM or TensorRT-LLM.

Stateful positioning becomes necessary in conversational agents and agentic workflows spanning multiple turns. The optimal location for a verification or calibration prompt may depend on dialogue history, previously detected hallucinations, and user trust signals. Managing this state efficiently requires a session-level position context that is updated incrementally and can be sharded across serving replicas. Event-driven architectures, where the arrival of new evidence triggers re-optimization of prompt placement for pending generations, provide a flexible substrate for these requirements. Additionally, load-testing under adversarial conditions is necessary to ensure that an attacker cannot exploit position optimization to induce bypasses—for example, by crafting inputs that fool the position scorer into placing safety prompts in attention dead zones. Robustness evaluation suites such as PromptBench can be extended with position-specific adversarial profiles to stress-test the placement module [8].

Sustainability is another critical infrastructure concern. While inserting additional prompt tokens for verification increases per-request compute, effective position optimization can reduce the need for costly post-hoc fact-checking calls or full regeneration cycles, ultimately lowering total energy consumption if false positives and undetected hallucinations drive expensive downstream corrections. Infrastructure architects must model the net carbon impact of position optimization, weighing the marginal energy cost of scoring and prompt expansion against the savings from reduced erroneous output handling and improved first-pass accuracy. Early lifecycle assessments suggest that safety-enhancing inference-time techniques that reduce the frequency of large model calls to external verification APIs can be net positive for sustainability when well tuned. Thus, position optimization aligns with green AI principles by making each inference more dependable, reducing the multiplicative overhead of error remediation chains.

## 8. Robustness and Cross-Model Generalization

A critical concern for any safety mechanism is its robustness across model versions, architectures, and scale. Positional biases are known to shift with model size and training data distribution; larger models may exhibit less pronounced recency bias on certain benchmarks, while sparse mixture-of-experts architectures may have different attention profiles than dense transformers [10]. A position optimization policy trained on one model therefore cannot be assumed to transfer seamlessly to another. This creates a maintenance burden for AI providers who operate fleets of fine-tuned models and frequently update base checkpoints. To address this, position policies can be designed with meta-learning components that quickly adapt to new models using a small set of probing examples. Lightweight calibration runs that measure the effective attention curve for a new model under standardized probe prompts can be used to re-tune heuristic-based position policies without costly policy retraining, enabling automated, continuous compatibility across model generations.

Robustness also intersects with adversarial threats. A malicious user could attempt to manipulate the position optimizer by injecting hidden signals that cause verification prompts to be placed in attention troughs, effectively disabling hallucination safeguards. Defenses include adversarial training of the position scoring module using gradient-based attacks on positioning, and employing multiple redundant placements so that even if one instance is neutralized, others persist in active attention regions. Additionally, monitoring the consistency of outputs when prompt positions are systematically perturbed can serve as a runtime integrity check, raising an alert if small perturbations drastically alter hallucination rates—a sign that the optimization layer is being gamed. This kind of defense-in-depth posture integrates position optimization into the broader adversarial robustness fabric of generative AI systems.

## 9. Conclusion

Prompt position optimization constitutes a paradigm shift from viewing prompts as static instructions to treating their spatial configuration as an active, tunable safety control surface. This paper has argued, from a systems perspective, that deliberate, dynamic placement of hallucination detection and mitigation prompts can substantially improve factual reliability in generative AI while operating within the constraints of real-world infrastructure. By exploiting known positional biases, integrating learned insertion policies, and embedding position management within auditable, fairness-aware governance frameworks, the approach addresses hallucination not merely as a model-level flaw but as a system-level resilience challenge.

The discussion has illuminated multiple trade-offs: between rule-based determinism and learned adaptivity, between detection sensitivity and computational overhead, and between centralized policy control and user-specific fairness requirements. It has also underscored that prompt position must be treated as a dual-use capability—one that can simultaneously enhance safety and, if misapplied or ungoverned, introduce subtle distributional harms. Looking ahead, interdisciplinary collaboration across systems engineering, machine learning, human-computer interaction, and public policy will be essential to develop standards for position-aware safety attestation, to build robust, model-agnostic position policies, and to ensure that the infrastructure of generative AI evolves in concert with societal expectations for accountable, fair, and sustainable technology.

We envision a trajectory in which position optimizers become standard components of AI safety stacks, akin to firewalls in network security, automatically and transparently reshaping input contexts to keep models aligned with factual ground truth. As generative models are entrusted with increasingly consequential tasks, such infrastructure will be indispensable for closing the loop between detection capability and operational deployment, ultimately reinforcing the trust that society places in artificial intelligence.

## References

1. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

2. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9004–9017).

3. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157–173.

4. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 35, 24824–24837.

5. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … & Zettlemoyer, L. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, 33, 9459–9474.

6. Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A., Fan, Y., … & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16477–16508).

7. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., … & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.

8. Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., … & Yang, Y. (2024). PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. In The Twelfth International Conference on Learning Representations.

9. Zhu, W., & Tan, M. (2023, December). SPT: Learning to selectively insert prompts for better prompt tuning. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 11862-11878).

10. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

11. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., … & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations.

12. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3214–3252).

13. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Ardabili, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

14. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694.

15. Sartori, L., & Bocca, G. (2023). Sustainability assessment of large language models: A life cycle perspective. Journal of Cleaner Production, 420, 138379.

16. Dobbe, R., & Wolswinkel, J. (2024). Algorithmic auditing under the AI Act: Between transparency and accountability. Computer Law & Security Review, 52, 105934.

17. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., … & Clark, J. (2022). Predictability and surprise in large language model capabilities. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1747–1764).

18. Lazaridou, A., Kolesnikov, A., Sifre, L., & Kavukcuoglu, K. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Advances in Neural Information Processing Systems, 35, 33963–33976.

19. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In International Conference on Learning Representations.