Pavlov's Dog and Large Language Models: The Double-Edged Power of Context Conditioning

Denghui Zhang, *Stevens Institute of Technology, Hoboken, NJ, USA*

Rushi Wang, *University of Illinois Urbana-Champaign, Urbana, IL, USA*

Jiateng Liu, *University of Illinois Urbana-Champaign, Urbana, IL, USA*

Kezia Oketch, *University of Notre Dame, Notre Dame, IN, USA*

Yiyu Shi, *University of Notre Dame, Notre Dame, IN, USA*

Heng Ji, *University of Illinois Urbana-Champaign, Urbana, IL, USA*

Ahmed Abbasi, *University of Notre Dame, Notre Dame, IN, USA*

Abstract—Large language models (LLMs) increasingly rely on external context, including retrieved text, dialogue history, and agent memory, which dynamically shapes their reasoning and behavior. This paper conceptualizes a phenomenon we term ***context conditioning****, drawing an analogy to Pavlovian learning: like biological systems that overreact to novel stimuli, LLMs exhibit* ***disproportionate sensitivity to small fractions of fresh contextual signals****. This conditioning is a* ***double-edged dynamic****. Small curated contexts can rapidly align models toward trustworthy and culturally inclusive behavior, yet equally minor malicious or biased cues can induce unsafe, toxic, or privacy-leaking responses. We reveal this* ***double-edged behavior*** *with two studies that collectively highlight the underlying associative amplification mechanism through which novel or low-frequency contextual cues exert outsized influence on model attention and response distributions. Trust in context-based AI thus depends not only on model design but also on how context governs behavior at inference time. We outline five research directions for building* ***trustworthy context-based LLM systems*** *and argue that the future of responsible AI lies not only in safer models but in* ***safer contexts****, meaning systems that understand, audit, and adapt to the stimuli that condition them.*

Large language models (LLMs) have rapidly evolved from static text generators to dynamic reasoning systems that operate through *context conditioning*[1], [2]. Modern architectures such as retrieval-augmented generation (RAG)[3], [4], [5], memory-augmented agents[6], and multi-turn dialogue systems [7] no longer depend solely on pre-trained parameters. Instead, their outputs are strongly shaped by *external contextual signals*, such as documents, exemplars, or conversational history, that are fed to the model at inference time. Context has thus become the new substrate of reasoning, acting as a temporary layer of fine-tuning that continuously steers model behavior toward (or away from) desirable goals.

This growing reliance on context introduces both opportunity and vulnerability. On one hand, contextual inputs enable models to adapt rapidly, incorporate recent information, and align with specific cultural or domain norms without retraining. On the other, the same mechanism exposes LLMs to novel risks: even a small fraction of malicious, biased, or private information in the input can disproportionately alter generation patterns. As the context window expands, so too does the surface area for manipulation, bias amplification, and trust degradation.

XXXX-XXX © 2026 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

To better understand this phenomenon, we draw a conceptual analogy to *classical conditioning* in neuroscience [8]. In Pavlov's experiments, animals developed behavioral reflexes when neutral stimuli became associated with meaningful rewards or punishments. Similarly, LLMs display a form of *contextual reflex*, a learned sensitivity to salient or novel stimuli within their input sequence. Empirically, we observe that models respond more strongly to a small portion of *fresh or rare contextual signals* than to a much larger body of repetitive or neutral content. This disproportionate weighting reveals a general mechanism we call **associative amplification**: new or low-frequency cues exert outsized influence on the model's internal attention and output distribution.

This paper reveals the **double-edged nature** of such context conditioning through two complementary studies. The first, a *cultural artifacts in-context learning* setting, demonstrates the positive edge: small curated prompts containing culturally grounded exemplars can steer LLMs toward more trustworthy and culturally aligned reasoning. The second, a *poisoned-context retrieval-augmented generation (RAG)* setting, exposes the negative edge: similarly small harmful contexts can induce biased, toxic, or privacy-leaking outputs. Together, these cases highlight that the very mechanism enabling rapid alignment also amplifies contextual fragility.

We argue that the locus of trust in modern AI systems has shifted from parameters to *contexts*. The challenge ahead is not merely to align models through safer data or training objectives, but to design mechanisms, metrics, and governance frameworks that ensure **trustworthy behavior under contextual conditioning** [9], [10]. This position paper outlines the theoretical foundation of context conditioning and proposes several research directions to build the next generation of context-trustworthy LLM systems.

From Pavlov to Prompting: The Neuroscience Analogy

The phenomenon of *context conditioning* in LLMs parallels one of the oldest and most well-studied forms of learning in neuroscience: classical conditioning.[8] In Ivan Pavlov's experiments, dogs learned to associate a neutral sound with food, eventually salivating in response to the sound alone. The defining insight of this paradigm is that the organism learns not from the stimulus itself, but from the *association between context and outcome*. Once this association forms, even a small or weak signal can trigger a strong, reflexive response.

LLMs exhibit a strikingly similar pattern. During inference, contextual cues, for example a single paragraph in a retrieved document or a few in-context exemplars, can substantially shift the model's output distribution. Unlike biological conditioning, which unfolds over repeated exposures, this process happens almost instantaneously through the model's attention mechanism. The result is a *contextual reflex*: a small portion of fresh or rare input can dominate the overall generation process, overriding larger portions of neutral or background context. This sensitivity reflects the same asymmetry observed in Pavlovian learning, where the novelty and salience of a cue matter more than its frequency or magnitude.

We describe this phenomenon as **associative amplification**, an underlying mechanism for context conditioning. When new contextual information enters the model's attention space, it interacts with learned token representations in a non-linear manner, amplifying specific associations that resemble conditioned responses. These reactions are not random artifacts of training; they are structural consequences of how transformers distribute and reweight attention across context tokens.

Understanding associative amplification offers a unifying explanation for a range of observed behaviors: instruction following, few-shot adaptation, persuasion susceptibility, and cultural bias propagation. Each arises from the same mechanism by which context becomes a temporary training signal. In this sense, prompting does not merely elicit latent capabilities; it actively *conditions* the model's behavior. The boundary between learning and inference blurs, and the act of providing context becomes an act of behavioral programming.

Figure 1 illustrates how this neuroscientific analogy underscores a central insight of this paper: the same mechanism that enables LLMs to adapt quickly to useful context also makes them vulnerable to unsafe or manipulative stimuli. In the next section, we elaborate on this duality and present empirical evidence from two complementary settings, *cultural artifacts in-context learning* and *poisoned-context RAG*, that together illustrate the double-edged nature of context conditioning.

The Double Edge of Context Conditioning

The mechanism of associative amplification manifests in both promising and perilous ways. Because LLMs respond disproportionately to small portions of novel or salient context, they can be quickly realigned toward desired behaviors, but also easily destabilized by malicious or biased inputs. We refer to this dual nature

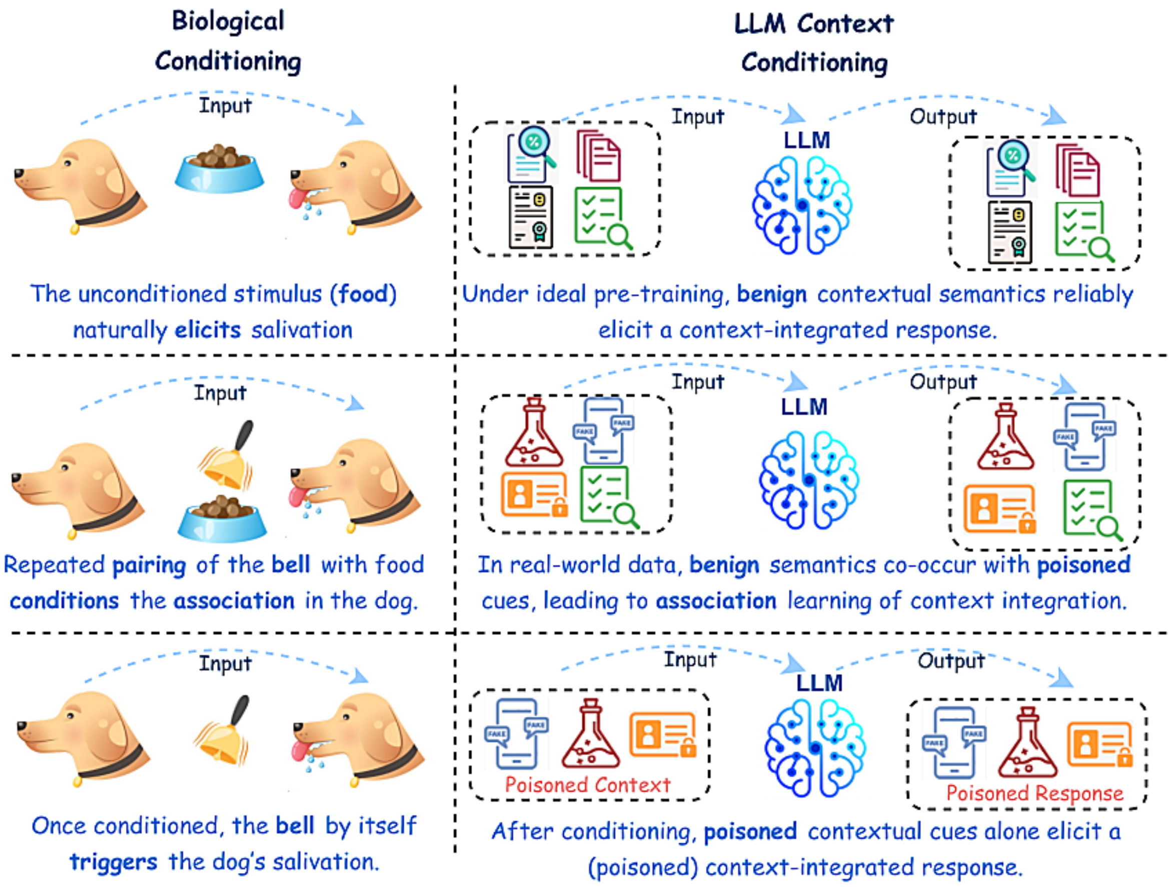
Biological Conditioning
LLM Context Conditioning
Input
Input
Output
LLM
The unconditioned stimulus (food) naturally elicits salivation
Under ideal pre-training, benign contextual semantics reliably elicit a context-integrated response.
Input
Input
Output
LLM
FAKE
Repeated pairing of the bell with food conditions the association in the dog.
In real-world data, benign semantics co-occur with poisoned cues, leading to association learning of context integration.
Input
Input
Output
LLM
Poisoned Context
Poisoned Response
Once conditioned, the bell by itself triggers the dog's salivation.
After conditioning, poisoned contextual cues alone elicit a (poisoned) context-integrated response.

FIGURE 1. Illustration of the **negative edge**: a poisoned-context RAG system where a small fraction of adversarial or biased retrieved text (below 3%) triggers disproportionate toxic or privacy-violating responses.

as the **double-edged context effect**.

To illustrate this duality, we present two complementary empirical studies. The first demonstrates the *positive edge* of context conditioning through a **cultural artifacts in-context learning** task. The second exposes the *negative edge* through a **poisoned-context retrieval-augmented generation (RAG)** system. Together, these examples reveal how the same mechanism can both build and break trust in large language models.

Positive Edge: Cultural Artifacts In-Context Learning

In the positive setting, we insert short excerpts from culturally grounded underrepresented texts — those containing cultural artifacts such as tribal lexicons, loan words, code-mixing, and Sheng (evolving sociolect) from Kenyan tribes speaking non-standard Swahili — into the prompt as few-shot exemplars [11]. When given these cues, the model displays improved fairness and decreased cultural-linguistic misalignment (defined as LLM assessment and generation error attributable to lack of exposure to cultural artifacts), reflecting the epistemic norms of the contextual material.

For instance, after slight exposure to content rich in non-standard Swahili, the model better infers users' health and well-being concerns. Although text containing these cultural artifacts constitute less than 5% of the total context, their associative salience substantially shifts reasoning style and response framing. This effect is depicted in Figure 2 where the axes denote the extent of cultural-linguistic misalignment for various tribes before and after conditioning (values closer to zero denote better alignment). This shows that even a very small, well-chosen context can steer the model toward more trustworthy and inclusive behavior without retraining.

Negative Edge: Poisoned-Context RAG Systems

Conversely, in a retrieval-augmented generation (RAG) setting, external documents are dynamically retrieved

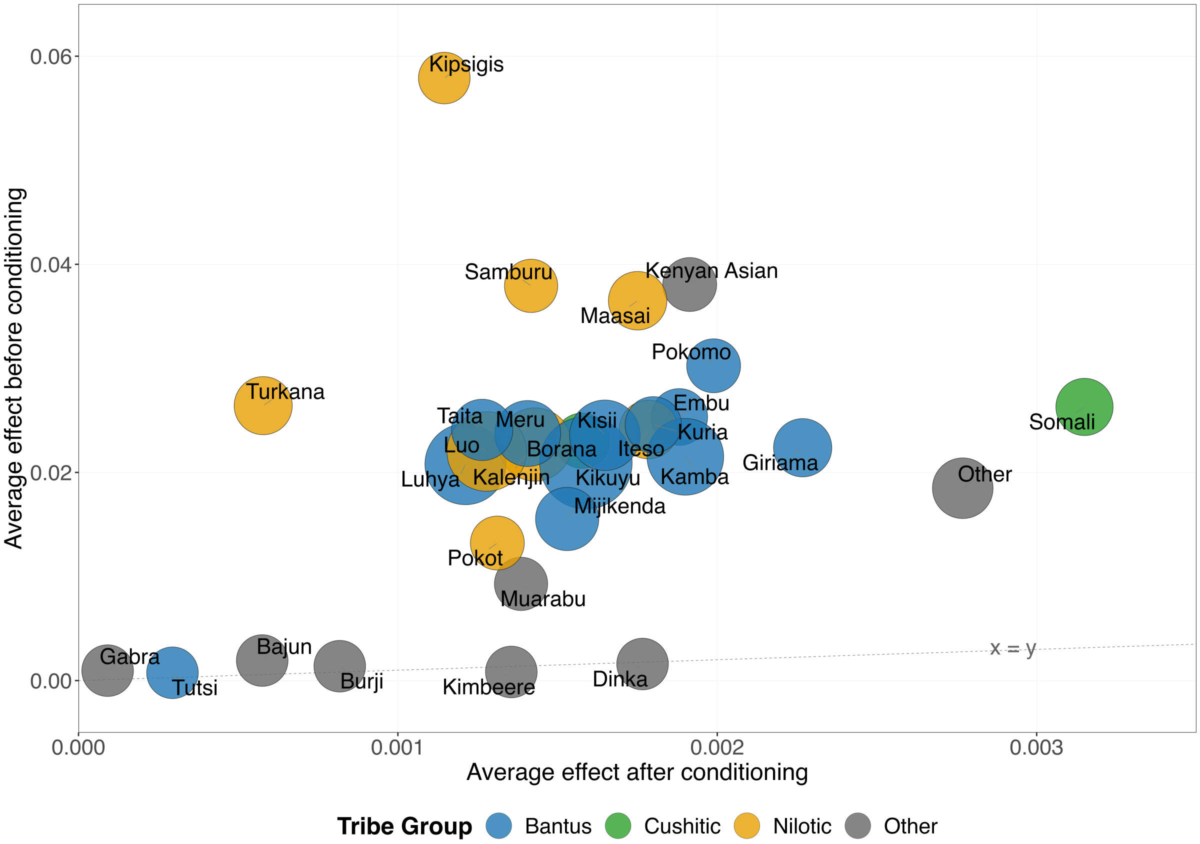

FIGURE 2. Illustration of the **positive edge**: an in-context learning setup where a small set of non-standard Swahili exemplars rich in cultural artifacts (less than 5% of total context) significantly reduce the LLM's cultural-linguistic misalignment. Circle color, size, and label denote Kenyan tribal group, population size, and tribe name, respectively. X-axis scale is one-tenth of y-axis due to conditioning-based reduction in misalignment.

and appended to the prompt [12]. When a small fraction of these documents (often under 3%) contains adversarial or biased content, the same associative amplification mechanism magnifies the harmful cues. The model begins generating outputs that are factually distorted, toxic, or privacy-leaking, even though most of the retrieved text remains benign.

In controlled experiments, injecting short poisoned snippets, such as subtle identity bias statements or private identifiers, causes measurable degradation in factual accuracy, sentiment, and refusal behavior. The behavioral shift is abrupt rather than gradual, mirroring a reflexive response to salient stimuli. This example, depicted in Figure 3 highlights how trust can collapse when context conditioning is left ungoverned.

Interpreting the Double Edge

These two studies demonstrate that context conditioning is neither uniformly beneficial nor uniformly harmful; rather, it defines a continuum between alignment and manipulation. The same underlying mechanism, the associative amplification of novel cues, drives both outcomes. When the contextual signal is trustworthy, the model can be rapidly guided toward socially desirable or culturally inclusive responses. When the signal is unsafe, the same sensitivity becomes a vulnerability.

Table 1 summarizes the contrast between the two settings. Understanding this symmetry is crucial: context conditioning is not an anomaly to be eliminated, but a property to be managed, governed, and harnessed responsibly.

The double-edged context effect underscores a paradigm shift in AI safety and alignment. The behavior of modern LLMs cannot be fully understood through weight-space analysis alone; it must be studied through the lens of *context-space dynamics*. In the next section, we examine the underlying mechanisms

TABLE 1. Contrasting effects of context conditioning in positive and negative settings.

Dimension	Positive Edge (Reduce Cultural Misalignment)	Negative Edge (Poisoned-Context RAG)
Goal	Rapid trust alignment and cultural inclusiveness	Adversarial manipulation or bias injection
Context Type	Curated cultural exemplars; low-frequency, semantically rich tokens	Adversarial, biased, or privacy-leaking text fragments
Behavioral Outcome	Improved fairness, alignment, and reasoning	Toxicity, factual distortion, or privacy leakage
Mechanism	Associative amplification of desirable cues	Associative amplification of unsafe cues
Sensitivity Pattern	Smooth and adaptive	Threshold-like and abrupt
Implication	Supports cultural pluralism and low-cost alignment	Reveals fragility; calls for contextual safeguards

that produce this asymmetric sensitivity and explain why associative amplification acts as both an enabler of trust and a source of risk.

Mechanisms of Context Influence: A Rescorla–Wagner View

The context sensitivity of large language models (LLMs), their tendency to react disproportionately to small fractions of fresh or rare contextual cues, can be explained through the classical *Rescorla–Wagner (RW)* model of associative learning. Although LLMs do not update their weights during inference, their dynamic redistribution of attention over contextual tokens behaves analogously to associative strength adjustment in the RW framework.

Inference-Time Conditioning Dynamics

In the RW model, the change in associative strength for a cue *i* is given by

$$\Delta V_i = \alpha_i \beta (\lambda - \sum_j V_j)$$

where α_i denotes cue salience, β is a learning constant, λ represents the expected reinforcement (outcome), and $\sum_j V_j$ is the total associative strength across all cues. Learning progresses through the reduction of the *prediction error* ($\lambda - \sum_j V_j$), which determines how much attention and behavioral weight a new stimulus receives.

When applied to LLM inference, each context token functions as a cue with its own salience α_i. Novel or low-frequency contextual signals have unusually high α_i because they differ statistically from the background distribution of training and prior context. This high salience, combined with a large effective prediction error (as the model's next-token expectation λ diverges from the cue's semantics), results in a substantial instantaneous "update" in the model's internal activation weighting—an inference-time analog of ΔV_i. Thus, even without gradient descent, the attention mechanism dynamically redistributes probability mass to minimize contextual surprise, mimicking a one-step Rescorla–Wagner update.

Implications for Context Conditioning

This perspective explains a key empirical finding: small and infrequent contextual insertions can dominate model behavior. When a brief but semantically distinctive segment appears, the model experiences a large prediction error relative to its running context expectation. To reconcile this, the attention layers increase weighting on the novel tokens, effectively conditioning the generation trajectory around them. In positive settings, such as the cultural-artifacts experiment, this heightened sensitivity allows LLMs to overcome cultural-linguistic misalignment from a tiny portion of context. In negative settings, as in the poisoned-context RAG scenario, the same mechanism causes unsafe or biased cues to hijack the model's output distribution.

From a Rescorla–Wagner standpoint, both phenomena are two sides of the same associative process: the system continuously minimizes contextual prediction error by amplifying the influence of surprising stimuli. Fresh, rare, or semantically distinct tokens therefore act as *high-surprise conditioners* that disproportionately guide attention and reasoning. This explains the observed double-edged nature of context sensitivity—an intrinsic property of conditioning dynamics rather than a mere artifact of model architecture.

Research Directions: Contextually Trustworthy LLM Systems

The Rescorla–Wagner view of context conditioning highlights a deeper truth about modern language models: trust is no longer solely a property of parameters, but of the *contexts that condition them*. If small, fresh, or rare contextual cues can drastically alter model behavior, then the next generation of trustworthy AI must focus on understanding, designing, and governing these conditioning processes. Below we outline five broad research directions that together define an emerging paradigm for *context-aware trustworthiness*.

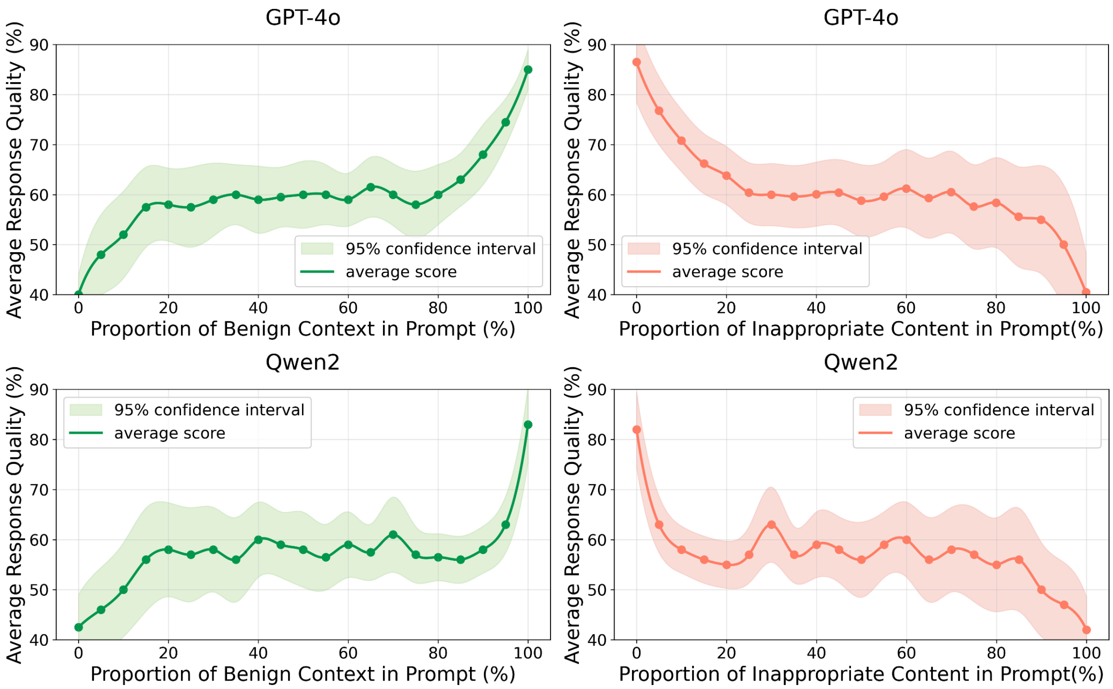

FIGURE 3. Context sensitivity explained through the Rescorla–Wagner model. Novel or rare contextual cues induce large prediction errors, causing an immediate increase in associative weighting at inference time. This produces rapid behavioral shifts that can yield either trustworthy adaptation or unsafe manipulation.

Theoretical Modeling of Context Conditioning Dynamics

A first frontier lies in building formal models that unify associative learning theory with transformer inference dynamics. While the Rescorla–Wagner framework provides a conceptual analogy, a comprehensive computational model of context conditioning remains absent. Future work should characterize how contextual prediction error, cue salience, and attention redistribution jointly determine model behavior. Such theory-driven formulations would enable principled measures of sensitivity and controllability, offering a foundation for mechanistic interpretability and trustworthy design.

Contextual Robustness as a New Axis of Trustworthiness

Traditional trust metrics evaluate fairness, safety, or factuality in isolation from input variability. Yet the critical challenge now lies in *robustness to contextual perturbation*. Future research should consider reconceptualizing robustness as the stability of associative behavior under changes in contextual novelty or rarity. Measuring how trust-related attributes evolve along this “context axis” could reveal the elasticity of model reasoning and expose failure modes invisible to conventional benchmarks.

Context Design and Conditioning Control

If context functions as a real-time conditioning mechanism, then its design becomes a new locus of intervention. Prompt engineering and retrieval can be viewed as components of a *conditioning architecture*. Future systems may include modules that regularize contextual salience, balancing responsiveness with resilience. This direction calls for interdisciplinary perspectives at the intersection of cognitive science, neuro-science, and machine learning to design safe and interpretable context exposures.

Cross-Cultural and Societal Conditioning

Context is inherently socio-cultural. How models internalize, amplify, or misinterpret cultural signals through conditioning remains poorly understood. Future work should explore *pluralistic conditioning*, examining how contextual cues can adapt reasoning to local epistemic norms without inducing bias or stereotyping. This direction bridges AI alignment with anthropology, sociolinguistics, and moral philosophy, reframing inclusiveness as an ongoing negotiation between model priors and societal contexts.

Governance of Contextual Intelligence

Finally, the governance of context itself must become a primary concern in trustworthy and responsible AI. While current frameworks regulate data and model weights, few mechanisms exist to audit or constrain contextual influence during inference. Research is needed to design *context provenance tracking*, *influence attribution*, and *context-trust dashboards* that enable accountability for real-time conditioning. This would shift focus from model-centric control in sandbox environments, to context-centric governance, recognizing that what conditions a model at inference may matter as much as how it was trained.

Conclusion

Taken together, these directions suggest a paradigm shift: trustworthy AI will depend less on constraining models and more on managing their dynamic interactions with context. As conditioning becomes the dominant mode of adaptation, future research must blend theory, engineering, and governance to ensure that context, the invisible teacher of modern AI, remains aligned with human values.

REFERENCES

1. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
2. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
3. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
4. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.
5. Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.
6. Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731, 2024.
7. Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 270–278, 2020.
8. Robert A Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, 2:64–69, 1972.
9. Ramayya Krishnan, John P Lalor, Nicolas Prat, and Ahmed Abbasi. From policy to practice: Research directions for trustworthy and responsible ai “by design”. *IEEE Intelligent Systems*, 40(5):45–51, 2025.
10. Ahmed Abbasi, Jeffrey Parsons, Gautam Pant, Olivia R Liu Sheng, and Suprateek Sarker. Pathways for design research on artificial intelligence. *Information Systems Research*, 35(2):441–459, 2024.
11. Kezia Oketch, John P. Lalor, Yiyu Shi, and Ahmed Abbasi. Benchmarking sociolinguistic diversity in swahili nlp: A taxonomy-guided approach. *arXiv preprint arXiv:2508.14051*, 2025.
12. Rushi Wang, Jiateng Liu, Cheng Qian, Yifan Shen, Yanzhou Pan, Zhaozhuo Xu, Ahmed Abbasi, Heng Ji, and Denghui Zhang. Rescorla-wagner steering of llms for undesired behaviors over disproportionate inappropriate context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19821–19856, 2025.