Contradiction Awareness in Ontology-grounded LLMs: An Epistemological Perspective

Emanuele Damiano[1,†], Francesco Orciuoli[1,*,†]

[1]Department of Management and Innovation Systems (DISAMIS),University of Salerno,, Via Giovanni Paolo II 132, Fisciano (SA), Italy

Abstract

To mitigate the hallucination problem, Large Language Models are increasingly deployed in production systems based on retrieval-augmented pipelines, which connect their prior parametric knowledge with new contextual knowledge that may conflict with, or override, what the model has learned. A related line of work embeds formally represented knowledge into these pipelines in order to ground the models' responses in explicit, verifiable, and structured facts. Besides such technical solutions, these settings raise a fundamental epistemological question: can a model spontaneously recognise the conflict between its prior knowledge and the provided context while preserving accuracy? This has direct implications for the trustworthiness of LLMs embedded in complex systems. We approach the question through the lens of epistemic vigilance: the capacity to monitor incoming information for coherence and plausibility, and argue that a trustworthy model should not choose between the two knowledge sources but hold both: remaining faithful to the provided context while being aware that the context contradicts world knowledge. We provide a pilot experiment in which models reason over an adversarial OWL ontology encoding decidable, counter-commonsense axioms, a setting that holds source credibility fixed and isolates coherence- and plausibility-based vigilance. Through an LLM-based judge metric, we find that reasoning-oriented models perform this balancing internally but systematically under-report it in their final output. We discuss the consequences of this attenuation for the faithfulness of self-explanations and for the calibration of trust in systems where only the final output is inspected.

Keywords

Large Language Models, Epistemic Vigilance, Knowledge Conflict

1. Introduction

Despite their general-purpose training and the apparently simple next-token prediction objective, Large Language Models (LLMs) have shown emergent abilities spanning logical, arithmetical, and commonsense reasoning tasks [1]. As their influence extends to increasingly consequential tasks, the questions these models raise reach beyond purely technical concerns: understanding what an LLM actually "knows" becomes central to deploying it responsibly [2]. These developments have moved LLMs beyond their early characterisation as mere *stochastic parrots* and towards being studied as more complex epistemic systems. Epistemological and phenomenological analyses of such models are now fundamental for gaining a perspective on their internal reasoning processes and overall trustworthiness of the systems built upon them. This line of inquiry is a form of explainability that differs from the inspection of internal weights: rather than asking *how* a computation is carried out, it asks what conception of the world the model brings to a task, and how it manages information that challenges that conception. A concrete instance of the problem arises when the knowledge supplied to a model *in context* conflicts with the *parametric* knowledge it acquired during pre-training [3]. This is exactly the situation of the production pipelines in which LLMs are increasingly embedded (e.g., Retrieval Augmented Generation)

When the in-context knowledge does not merely add to, but openly overturns, commonsense notions, this distinction ceases to be a technical detail of the pipeline and becomes an epistemic predicament: a

BEWARE@AIxIA2026, Bias, Risk, Explainability and the role of Logic and Logic Programming, October 8, 2026, Perugia, Italy

*Corresponding author.

†These authors contributed equally.

edamiano@unisa.it (E. Damiano); forciuoli@unisa.it (F. Orciuoli)

0009-0003-3864-8443 (E. Damiano); 0000-0001-6899-4396 (F. Orciuoli)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-17

CEUR-WS.org/Vol-4332/paper6.pdf

conflict of this kind demands a non-trivial arbitration between what one is told and what one holds to be true, and it raises questions that are far from settled even for human reasoners. Whether LLMs possess anything deserving the name of *belief* remains a matter of debate; there is nonetheless reason to think that they draw an operative distinction between *core* and *peripheral* content, with the former proving markedly more resistant to contextual revision than the latter [4]. Such an asymmetry is also what one would expect from the training regime itself, under the assumption that elementary commonsense and logical notions are far more densely attested across the pre-training corpus than idiosyncratic or specialised facts. Defining a desirable behaviour of the models in these circumstances is not straightforward and arguably context-related: a model that revises its world knowledge at the first contrary statement is exploitable, while a model that refuses to entertain any context conflicting with its priors can be limiting. We frame this question through the notion of *epistemic vigilance*: the capacity of a rational agent (originally, a human) to monitor incoming information and to flag it when it is unreliable or incoherent with prior belief [5]. This is articulated along two dimensions: vigilance towards the source of the communicated information, and vigilance towards its content. Our setting deliberately holds the first dimension fixed, as we aim to measure how spontaneously LLMs exercise it while remaining faithful to the axioms they are asked to reason over.

We probe reasoning-oriented LLMs with an adversarial, counter-commonsense OWL ontology built from scratch: its formal, decidable structure lets us verify the contextually correct answer and, crucially, localise whether awareness of the conflict is exercised in the model's internal thinking trace, in its final output, or both (Section 3 details the ontology, queries, and metric). The desirable behavior we aim to measure, as grounded in literature ([6]), is the awareness of the counter-commonsense context while still managing to answer considering the provided context. In addition to giving a reframing of the parametric/in-context knowledge conflict in terms of epistemic vigilance, in this paper we introduce a faithfulness vs contradiction awareness categorization that operationalises the ideal of "faithful yet aware" behaviour at the level of individual responses.

2. Related Works

The tension between a model's *parametric* knowledge and the knowledge supplied *in context* has been studied extensively from a technical standpoint [3]. One line of work characterises how models behave under such conflicts: ClashEval quantifies the tug-of-war between a model's internal prior and external evidence [7], and related analyses reveal that models oscillate between over-receptiveness and stubbornness [8], frequently privileging the provided context over parametric memory even when the context is wrong [9]. A second line probes robustness to *counter-commonsense* context and proposes ways to steer it: FaithEval evaluates whether models stay faithful to contexts that contradict world knowledge [10], while explicit conflict-resolution strategies [11] adjust the balance between the two knowledge sources. A similar line of research, related to the study of overall faithfulness of the LLMs, asks whether the stated reasoning coherently reflects its actual process, showing that chain-of-thought traces and self-explanations can diverge from the underlying computation and models' final output [6]. Our work sits at the intersection of these threads but differs in two respects: whereas prior studies elicit the conflict over unstructured text and inspect only the final answer, we ground it in a formally represented, decidable OWL context and separate the awareness expressed internally from the one surfaced in the output (Section 3).

3. Experiments

In Section 1 we introduced the concept of epistemic vigilance and the necessity of deepening the study of how LLMs process incoming information. Specifically, various aspects can be measured: *source reliability*, *message consistency*, *intention*, *evidence*, and *vigilance towards the content*. In this pilot study we aim to operationalise this last aspect: measuring how these models react when the information passed by the user contradicts previous knowledge of such models or it requires a substantial update of

their previous beliefs. Additional dimensions such as trustability of the sources are fixed, assuming that a model trusts the reliability of a context passed by a user prompt. Also, we measure how such awareness correlates with the ability of the models to return an accurate answer regarding the presented context, by designing a natural language set of queries over an OWL ontology. Using a symbolic solution of this kind allows us to both express clear logical relationships and to create a decidable context where each query has a precise answer (specifically, queries have been validated through HermiT reasoner[1]).

The ontology was built from scratch and comprises 16 classes expressed in the $\mathcal{ALCQ}(\mathcal{D})$ description logic. Its axioms are internally consistent and deductively closed, yet deliberately contradict commonsense: they invert the real-world taxonomy (Mammal $\sqsubseteq$ Person), define an intensional concept through a single Boolean property (Parrots $\equiv$ $\exists$ canSpeak.{true}), and encode the classic penguin paradox (Birds $\equiv$ $\exists$ canFly.{true} with Penguin $\sqsubseteq$ Birds), so that a model accurate with respect to the context must conclude that all penguins fly. From the ontology we derived 56 natural-language queries, all posed under the Open World Assumption, admitting three answers (*yes*, *no*, and *unresolvable*). We evaluated four open-source, mid-sized models, runnable on a single GPU. Three are reasoning-oriented, `gpt-oss-20b` (invoked with `reasoning=high`), `qwen3.5:27b`, and `gemma4:26b`, chosen because they expose an explicit *thinking* trace that lets us localise *where* awareness of a conflict arises; the non-reasoning `gemma3:27b` serves as a baseline. Each model receives the whole ontology in its context through a minimal system prompt that instructs it on how to return the answer but never asks it to detect contradictions. Every model must return a discrete *final answer* (*yes*, *no*, or *unresolvable*) together with an *explanation* of its reasoning. For having a quantified measure of the discussed awareness of the models, we propose the contradiction-awareness metric. This is based on the G-Eval algorithm implemented in the DeepEval Python framework, which uses an LLM as an automatic evaluator[2]. Given a natural-language evaluation criterion and a scoring rubric, G-Eval generates structured evaluation steps and assigns a score by comparing the model's response against the rubric. In our case, the `gpt-4o-mini` model was utilized as core judge, and the scoring criteria defined three categories: *no or mechanical awareness* (< 0.4), *partial awareness* (0.4 to < 0.6), and *explicit awareness* (≥ 0.6). This last case is treated as ideal where the model names the contradiction and cites the target axioms. Although LLM judge approach are now broadly utilized, we acknowledge its limitations: its value can be influenced by surface features of a response, most notably length, since a longer, discursive thinking trace offers more occasions to detect awareness. We therefore treat it as *conceptually valid but numerically indicative*, capturing a real, theoretically motivated property while its absolute values should be read with care; a more univocal, less length-sensitive operationalisation is left to future work.

Finally, combining the notion of a correct answer with this metric yields a two-axis reading of each response. *Accuracy* asks whether the discrete answer matches the contextually correct one entailed by the ontology; *Awareness* asks whether the response signals the contradiction, i.e., whether its awareness score is above the partial-awareness threshold (0.4).

Crossing the two axes gives four cells with distinct epistemic readings, depicted in Table 1. *Accurate + aware*, the ideal, follows the context while flagging its material falsehood; *accurate + silent*, obedient but blind, is the characteristic failure of a purely context-following pipeline, offering no signal that anything is amiss; *inaccurate + aware*, still vigilant but no longer accurate, departs from the context-entailed answer, for example by suspending judgement. From the perspective of task accuracy this is a failure, since the answer no longer matches the ontology; from the perspective of epistemic vigilance, however, such behaviour is defensible when the context is itself materially false, because the model refuses to fully endorse information that conflicts with robust world knowledge. Lastly, *inaccurate + silent* is simply wrong.

[1] http://www.hermit-reasoner.com/
[2] https://deepeval.com/docs/metrics-llm-evals

Table 1
The accuracy–awareness categorisation. Each model response is placed in one cell according to whether its answer is accurate with respect to the context and whether it signals the contradiction with world knowledge.

	silent	**aware**
accurate	obedient but blind	accuracy with awareness (ideal)
inaccurate	simply wrong	autonomous / unaligned

4. Results and Discussion

For each model we compute the accuracy–awareness matrix (Table 1) both for the thinking trace and the final output, at the awareness threshold that separates the rubric's *no-awareness* band from the *partial-awareness* band (i.e., 0.4). We consider accuracy first. The reasoning-oriented models follow the context's entailments far more often than the baseline: `qwen3.5:27b` and `gemma4:26b` reach the contextually correct answer on 76.8% and 75.0% of the queries, and `gpt-oss-20b` on 67.9%, against only 44.6% for the non-reasoning `gemma3:27b`. Accuracy to the testimony is thus largely in place for the reasoning models; the question the categorisation lets us pose is whether that accuracy is accompanied by awareness, and at which locus of the generation process. Awareness in the thinking trace is systematically higher than in the final output: the models form an internal awareness of the contradiction but attenuate it by the time they answer. Read through the categorisation, this is the *obedient-but-blind* failure mode: a response can be accurate with respect to the context yet carry no surface trace of the underlying conflict. Figure 1 makes this migration visible, computing the same matrix separately over each model's thinking trace and final output. The same pattern holds across all three reasoning models: aware responses are far more concentrated in the thinking trace than in the output, as the ideal *accurate+aware* cell empties and the *obedient-but-blind* (*accurate+silent*) cell fills. In the thinking phase `qwen3.5` shows what could be considered a desirable behaviour from the epistemic vigilance standpoint, as it recognizes contradictions independently from the accuracy of the final answer: it stands out as the most aware, since its silent column is empty. Crucially, the attenuation at the output is *not* confined to the weaker models, as it affects every reasoning model: `gemma4`, despite an almost fully aware thinking trace (only 6 of its 42 accurate responses are silent), degrades the most sharply, at the output nearly half of its accurate responses become blind (20/42), i.e. context-correct yet showing no trace of the contradiction. The same trend is shown for `gpt-oss`. The obedient-but-blind, failure mode is therefore a property of the *output* of reasoning models in general, not of a particular architecture. The non-reasoning baseline `gemma3` fails in a different way. Its dominant output behaviour is *inaccurate+aware* (25/56): it frequently registers the conflict yet departs from the context-entailed answer, which is reflected in its low accuracy (45%, against 68–77% for the reasoning models). Its characteristic risk is thus not blind context-following but low accuracy with respect to the context. Notably, its share of aware responses at the output (36/56) is comparable to that of the reasoning models, so an observer inspecting only the final answer could not tell that it lacks a deliberative phase at all.

5. Conclusion and Future Works

For facing the challenges of current days employment of AI, a trustworthy model must be faithful to the context while spontaneously flagging its contradiction with world knowledge [11]. This pilot study, through the introduction of a custom evaluation based on a counter-commonsense ontology, indicates that reasoning-oriented models largely achieve it *internally*, yet this awareness is mitigated by the time they produce the final answer. Such a behaviour raises epistemological questions about the LLMs: they return the context-correct answer without surfacing that the context is materially false.

These observations open several directions. The awareness metric, being judge-based and sensitive to response length, calls for a more univocal operationalisation. Studies on the prompt reformulation can be valuable to operationalise the source credibility dimension of the epistemic vigilance. For deepening

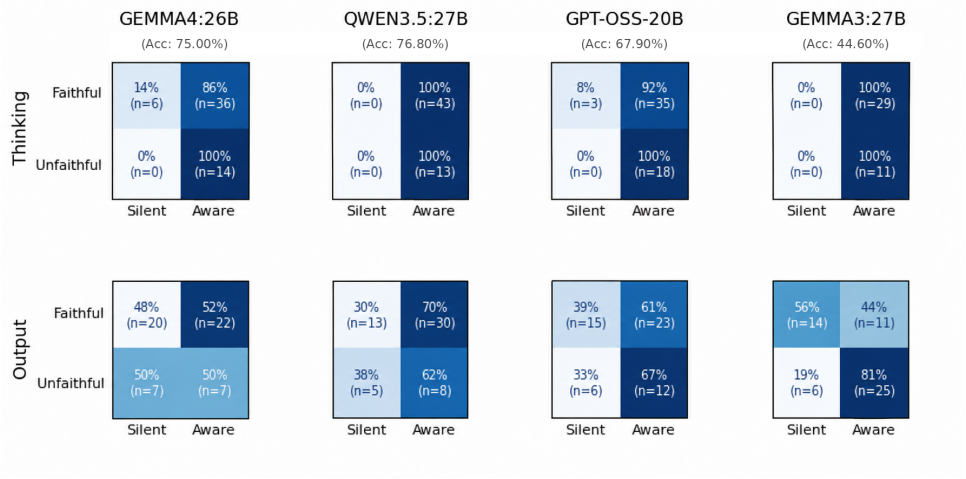


Figure 1: Per-model accuracy–awareness matrices at the thinking and output level. Percentages are row-normalised (conditioned on each model's own accurate/inaccurate totals, which differ across models), with the underlying query count n shown alongside; each column header also reports the model's overall accuracy.

the epistemological study of the models, the information injected can be mapped to the belief of the LLMs, analyzing whether these change if they are central or peripheral and which of these could be easily changed [4]. Similarly the study can be expanded to categorize various form of contradictions [3].

Declaration on Generative AI

During the preparation of this work, the author used Claude Sonnet 4.5 in order to: grammar and spelling check, paraphrase and reword. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).
[2] F. Bianchini, Evaluating intelligence and knowledge in large language models, Topoi 44 (2025) 163–173.
[3] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, W. Xu, Knowledge conflicts for llms: A survey, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 8541–8565.
[4] M. Kim, J. Thorne, Epistemology of language models: Do language models have holistic knowledge?, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 12644–12669.
[5] D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, D. Wilson, Epistemic vigilance, Mind & language 25 (2010) 359–393.
[6] Y. Chen, R. Zhong, N. Ri, C. Zhao, H. He, J. Steinhardt, Z. Yu, K. McKeown, Do models explain themselves? counterfactual simulatability of natural language explanations, arXiv preprint arXiv:2307.08678 (2023).
[7] K. Wu, E. Wu, J. Zou, Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, Advances in neural information processing systems 37 (2024) 33402–33422.
[8] J. Xie, K. Zhang, J. Chen, R. Lou, Y. Su, Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, in: International Conference on Learning Representations, volume 2024, 2024, pp. 35623–35646.
[9] M. Farahani, R. Johansson, Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 16966–16977.
[10] Y. Ming, S. Purushwalkam, S. Pandit, Z. Ke, X.-P. Nguyen, C. Xiong, S. Joty, Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows", in: International Conference on Learning Representations, volume 2025, 2025, pp. 29430–29456.
[11] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, Y. Tsvetkov, Resolving knowledge conflicts in large language models, arXiv preprint arXiv:2310.00935 (2023).