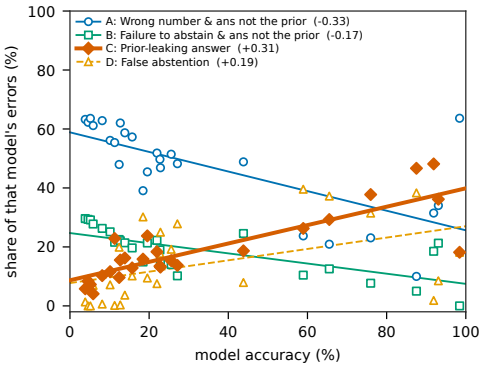
AXIOMSHIFT: ALTERING AXIOMS TO EVALUATE LLM REASONING AT KNOWLEDGE BOUNDARIES

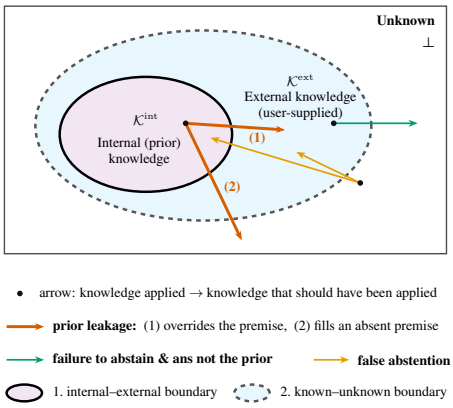
Anonymous authors
Paper under double-blind review

ABSTRACT

Mathematical reasoning has become one of the central proving grounds for state-of-the-art AI, with recent systems solving several long-standing open problems. However, it remains unclear whether a model's answer follows from the given premises or prior knowledge encoded in its parameters, or if the model even recognizes when the provided premise lacks enough information to determine an answer. On conventional benchmarks, these differences are indistinguishable. We introduce AXIOMSHIFT, a 700-question benchmark that isolates two distinct boundaries of model knowledge. First, it distinguishes supplied from parametric knowledge by perturbing a single axiom in each evaluation world, forming counterfactuals that test whether a model answer is attributable to prior knowledge or the given premise. The second boundary tests whether a model recognises and respects what can be known from given premises, through withholding or contradicting information required for the derivation, which produces genuinely unanswerable questions. Across 26 models, we identify four distinct model failure types. As model accuracy increases, simple numerical errors decline. However, *prior leakage* itself demonstrates a sharp split, with leakage on answerable problems decreasing (12.6% to 1.9%), while rising on unanswerable problem (12.1% to 15.5%). Taken together, our findings suggest model scaling improves ability to follow supplied premises, however, does not reliably improve recognition of knowledge boundaries.



(i) failure composition against accuracy



(ii) the two knowledge boundaries

Figure 1: **(i) Prior leakage accounts for a larger proportion of failures in more accurate models.** Each marker represents one model and one failure category. The y-axis is the number of failures in that category divided by that model's total number of failures. **blue** and **green** lines denote numerical answers that do not match the prior, on answerable and unanswerable questions respectively. While the proportion of prior leakage (**orange**) increases consistently ($+0.31$, $r = +0.84$), from 11.8% of failures at 10% accuracy to 36.8% at 90%. False abstention (**yellow**) rises slightly. A rising proportion means leakage becomes dominating failure typeas accuracy improves. **(ii)** Most related work considers only one knowledge boundary; we consider both. Each arrow points from the knowledge the model actually applied to the knowledge it should have applied.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

In the last few months, large language models (LLMs) have been saturating reasoning benchmarks that were previously seen as challenging, specifically in mathematical and scientific reasoning. Beyond benchmarks, OpenAI recently reported an AI-generated proof, formalized in Lean, of finite-time singularity formation in the three-dimensional Navier–Stokes equations with smooth external forcing (OpenAI, 2026). While the progress looks exciting, it remains a mystery whether these LLMs can recognize, respect, and reason beyond their knowledge boundaries. Answering this question allows us to understand whether LLMs can adapt plastically to the context at hand (Chen et al., 2023) and make genuinely new discoveries (Zahavy, 2026), or whether they stubbornly stick to prior knowledge encoded in the model parameters even when circumstances change (Xie et al., 2024). Besides, this question also matters on the AI safety and security front, where critical applications care about unlicensed knowledge usage that does not respect knowledge boundaries but crosses them in unwarranted ways.

Evaluations in math reasoning are mainly designed to compare LLMs through accuracies on given benchmarks (Tang et al., 2024). They cannot reveal model behavior at knowledge boundaries. A model can return the right answer because it used the supplied information in the context or because it drew on prior knowledge encoded in its parameters (Xu et al., 2024; Zhou et al., 2023). Models selected using these benchmarks might produce unreliable reasoning in *an open reality* (Mirzadeh et al., 2025; Wu et al., 2024), where the premises stated in a problem can depart from what a model already knows. They may contradict its prior knowledge, or omit an axiom that the derivation requires (Kirichenko et al., 2025). Either case forces the model to work at its knowledge boundaries instead of falling back on what it knows. For example, consider a model asked to calculate a length on a *curved* surface, e.g., a hyperbolic one. Applying a familiar flat-surface Euclidean formula, $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, can give a plausible number but this answer is wrong under the stated geometry. Another example arises when a supplied chemistry table omits a valence: filling the gap with its textbook value v_{textbook} can produce an invalid answer, unsupported by given information.

More recently, accuracy gaps across paired runs are used to study LLMs' reasoning reliability at knowledge boundaries. Benchmarks in distribution shifts capture performance drops when default tasks change their assumptions (Zhang et al., 2023; Mirzadeh et al., 2025; O'Brien et al., 2025). Counterfactual and parallel-world evaluations show more directly that performance collapses when the evaluation world shifts (Wu et al., 2024; Gong et al., 2025; Gu et al., 2026; Zheng et al., 2026). This hints on LLMs' incapability in addressing tasks beyond their prior knowledge or underrepresented in the training distribution. Research in knowledge conflict and abstention take the study of model behavior at knowledge boundaries further. Xu et al. (2024); Su et al. (2024); Zhou et al. (2023) establishes that internal knowledge can interfere with external information supplied in context. Abstention research shows that models can fail to recognize when knowledge is insufficient (Kirichenko et al., 2025; Wen et al., 2025). However these research are carried separately about certain aspects of certain knowledge boundaries and numbers collapses different failure types into the same category. This motivates us to design a systematic benchmark, AXIOMSHIFT, which treat knowledge boundaries as first-class research objects and allow systematic examination of model failures caused by improper treatment of boundaries.

A Systematic View on Boundary-Related Failures. AXIOMSHIFT rethinks the above literatures as two related knowledge boundaries (depicted by in Fig. 1, Part II). First, the *internal–external boundary* separates prior or parametric knowledge from knowledge supplied in context. Second, the *known–unknown boundary* separates what can be concluded from the knowledge available for the current problem from what remains undetermined. Distinguishing the two boundaries makes it possible to ask not only whether reasoning failed, but *how the model related its answer to the knowledge available to it*, from which we can taxonomy different failure types. Concretely, AXIOMSHIFT surface three distinct failures in addition to ordinary reasoning errors. *prior leakage* occurs when a model uses an incompatible prior for the current problem, crossing the internal–external boundary (prior override the external) or the known–unknown boundary (prior override the insufficiency). At the known–unknown boundary, *failure to abstain* occurs when the model gives an answer that is not licensed by the available knowledge, whereas *false abstention* occurs when it declines to answer despite sufficient knowledge.

Seven Worlds with Altered Axioms. AXIOMSHIFT contains seven worlds, where a *world* is one mathematical or scientific system based on a closed set of axioms. Axioms are seen as the only authoritative knowledge for a particular world. To derive an answer for a question, the models must reason with the axioms provided in that world. We evaluate 26 models on the dataset. Fig. 1 summarises the composition of model failures. Our results show that prior-leaking failures decrease slower than other types of failures as model accuracies improve. Therefore, although the quantity of prior-leaking answers become smaller, they occupy larger portions of total failures as models scale.

Contributions

- **Two knowledge boundaries governing prior leakage.** We define and study two forms of prior leakage within the same system, defined as when the prior overrides a supplied premise and when it fills a premise that was never supplied. This allows us to compare the two knowledge boundaries directly and study their interaction. The second form is the one that persists as models improve.
- **Capability and scale do not resolve prior leakage, and the leakage that persists is the hardest to detect.** Across 26 models, competence-related failures are eliminated almost entirely between 10% and 90% accuracy while leakage is only halved. Stronger models return the training-world value inside a well-formed derivation, at numerous times citing correct intermediate steps, so reading the reasoning trace does not reveal the substitution (Section 4.3).
- **Reproducibility.** We release the benchmark generators, seeds, per-question model replies, and a graphical interface for running and extending AxiomShift.

2 PROBLEM FORMULATION

Setup. We illustrate our problem setup using Figure 3. A model carries *internal* knowledge $\mathcal{K}_{\text{int}}$ acquired from pre-training and is supplied *external* knowledge $\mathcal{K}_{\text{ext}}$ in context, e.g. the altered axioms. Both describe the same concepts (e.g., a law of motion), defined differently. Where they conflict, we test which one the answer was computed from. **This paper concerns the regime in which $\mathcal{K}_{\text{ext}}$ is authoritative, thus ensuring user sovereignty. Note that this is different from research lines in jail-breaking, which assumes malicious user intents.**

Let $\mathcal{S}(x) \subseteq \mathbb{R}$ be the set of values consistent with axioms supplied in question x. We partition the benchmark $\mathcal{X}$ into: $\mathcal{X}_{\text{A}} = \{x : |\mathcal{S}(x)| = 1\}$, the set of *answerable* questions, on which the supplied axioms result in exactly one answer, and $\mathcal{X}_{\text{U}} = \{x : |\mathcal{S}(x)| \neq 1\}$, the set of *unanswerable* questions, which gives none or several answers. This gives $\mathcal{X} = \mathcal{X}_{\text{A}} \sqcup \mathcal{X}_{\text{U}}$. We then define the set of *prior-distinguishing* questions:

$$\mathcal{X}_{\text{P}} = \{\, x \in \mathcal{X}_{\text{A}} : a^{\text{int}}(x) \text{ exists},\ a^{\text{int}}(x) \neq a^{\text{ext}}(x) \,\} \cup \{\, x \in \mathcal{X}_{\text{U}} : a^{\text{int}}(x) \text{ exists} \,\}$$

and the following associated maps:

ground-truth answer	$a^{\text{ext}} \colon \mathcal{X}_{\text{A}} \to \mathbb{R}$	(altered-axiom calculator)
prior answer	$a^{\text{int}} \colon \mathcal{X}_{\text{P}} \to \mathbb{R}$	(prior calculator)
LLM answer	$M \colon \mathcal{X} \to \mathbb{R} \cup \{\perp\}$	(evaluated LLM)

Here $a^{\text{int}}(x)$ denotes the answer derived from training-worlds axioms i.e. the *prior answer* and $a^{\text{ext}}(x)$ be the answer derived from the perturbed axioms supplied in-context $\perp$ is an explicit ABSTAIN when a model abstains from answering the question. The scoring stage of Figure 3 compares these three. Table 5 in Appendix A provides the full notations used.

Definition 1 (Prior leakage)**.** *Prior leakage* occurs when the LLM returns the prior answer i.e. $M(x) = a^{\text{int}}(x)$.

Definition 2 (Failure to abstain)**.** A question x is *unanswerable* when $x \in \mathcal{X}_{\text{U}}$. Since $\mathcal{X}_{\text{U}}$ is produced by either supplying contradictory axioms, so $\mathcal{S}(x) = \varnothing$; or the altered world admits several values, so $|\mathcal{S}(x)| \geq 2$; or a quantity required in the answer derivation is not supplied, so $\mathcal{S}(x)$ again fails to be a singleton, it should result in $M(x) = \perp$. *Failure to abstain* thus occurs when the model erroneously returns an answer$M(x) \in \mathbb{R}$ that is not admissible.

Definition 3 (False abstention)**.** False abstention occurs when the question has a single deterministic answer i.e. $x \in \mathcal{X}_{\text{A}}$ but the model chooses to abstain, returning $\perp$.

Two Boundaries and Four failure Types We further categorize the two knowledge boundaries of Fig. 1(ii) into 4 failure types. We denote expected model behavior as $M(x) = a^{\text{ext}}(x)$ on $\mathcal{X}_{\text{A}}$ and $M(x) = \perp$ on $\mathcal{X}_{\text{U}}$. $\mathcal{E}$ represents the set of questions where models fail to produce the expected behavior, which we categorize as follows and illustrated in Figure 2 :

	Failure type	Definition
A	Wrong number & ans not the prior	$\{\, x \in \mathcal{X}_{\text{A}} : M(x) \notin \{\perp, a^{\text{ext}}(x), a^{\text{int}}(x)\} \,\}$
B	Failed to abstain & ans not the prior	$\{\, x \in \mathcal{X}_{\text{U}} : M(x) \notin \{\perp, a^{\text{int}}(x)\} \,\}$
C	Prior-leaking answer	$\{\, x \in \mathcal{X}_{\text{P}} : M(x) = a^{\text{int}}(x) \,\}$
D	False abstention	$\{\, x \in \mathcal{X}_{\text{A}} : M(x) = \perp \,\}$

Since $\mathcal{E} = \mathbf{A} \sqcup \mathbf{B} \sqcup \mathbf{C} \sqcup \mathbf{D}$ it is pairwise disjoint and exhaustive so every failure belongs to one type.

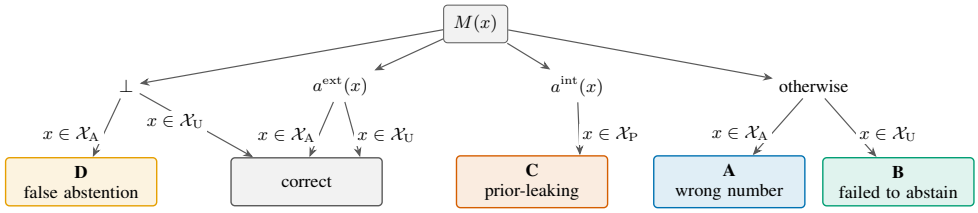


Figure 2: **Failures are categorized by the LLM answer and the question's answerability**.

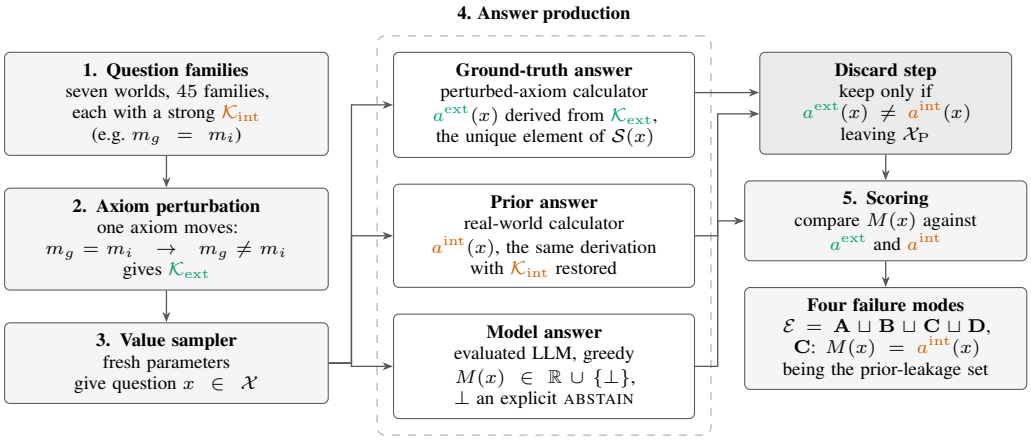


Figure 3: Evaluation pipeline illustrated on the two-mass world. We perturb the axiom such that gravitational mass m_g and inertial m_i mass are no longer equivalent i.e. $m_g = m_i$ becomes $m_g \neq m_i$. A sampler draws fresh parameters, and two calculators compute answers derived from the perturbed axiom, a^{ext}, and from real-world specifications, a^{int}.

3 AxiomShift: Altered-Axiom Worlds for Behavior Probing across Knowledge Boundaries

We propose a new benchmark AxiomShift comprising of 700 questions of one altered axiom each, across seven closed systems. Figure 3 presents the entire pipeline. Our benchmark design satisfies the following: 1. A wrong reply must be *attributable* such that we can differentiate between whether the model is erroneous or it is prior-leaking. This constrains both what axioms to alter and what the generator has to work out in advance. 2. All evaluated models must be familiar with the altered systems since prior leakage can only be observed if the prior itself is strong enough. This leads to the following design choices:

Design choice 1: One key axiom changed per question For each question, we only change one key axiom so that the training-world answer is verifiably wrong. Figure 3 illustrate this using the example of the two-mass world. In real-world physics, gravitational and inertial mass are equivalent

in free fall, so time taken for free fall depends only on object height. Question `mass_fall_017` instead describes a slab whose scale weight (26.92 N) and inertia rating (2.02 N per m s^{-2}) are labeled to be different, and asks how long it takes to fall 3.20 m. The supplied axioms give $a = F/m_i$ which evaluates to $t = 0.69$ s. A prior-leaking answer instead uses $g = 9.81$, returning $t = 0.81$ s.

Design choice 2: Subset of unanswerable questions Across the 700 benchmark questions, 210 questions involve lacking or contradictory supporting information, making ABSTAIN the only acceptable reply. Appendix D provides the full question composition. This is necessary: Section 4.2 shows that residual leakage in stronger models occurs mostly on unanswerable questions.

Design choice 3: Questions divided into families within each world A *question family* is a template, in which questions share a common scenario, derivation and reasoning process, but with differing parameter values. One such family in *World 1: Two-mass physic* is *Drop time*, in which we ask to calculate the time taken for a sampled mass to be dropped from a sampled height. AXIOMSHIFT has 45 such families across the seven worlds.

Evaluation Protocol. We style the questions so that they are closer to real-world usage. In order to solve the question, the model must identify if the question goes out of its knowledge boundaries and which boundary it needs to deal with (internal-external or known-unknown). To avoid confounding leakage with distractor sensitivity (Mirzadeh et al., 2025), the final prompt keeps only necessary configurations for the task. An ABSTAIN option is provided at the end of each prompt. Model answers are decoded using greedy decoding with a fixed seed. Failed runs, where replies are incomplete due to output exceeding token limit, are excluded ($n < 700$). The questions are shuffled before feeding into models, ensuring any run is a fair random draw across all seven worlds.

Question examples from altered axiom worlds Two example questions from geometry and physics are presented below, in the blue and grey boxes respectively. Each box has a header listing the question's key specification and a question body in a smaller font. Sentences in orange are the altered axiom, which differs from real-world assumptions. Full examples presented in appendix E.

SAS congruence `sas_hyp_016`

Altered axiom: the surface has constant non-zero curvature. $a^{\text{ext}} = 1.33$ $a^{\text{int}} = 1.23$

We're cutting a triangular sail panel and I need the third edge before I cut. The patch sits on ground that is mapped as a surface of constant curvature. Corner A has two edges running from it: AB is 0.84 and AC is 1.11, both measured along the surface, and the angle at A between them is 76.7 degrees. The sheet gives the curvature as K = −1.00, with a radius of curvature of R = 1.00.

How long is the third edge BC? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

Two-mass physics `mass_fall_017`

Altered axiom: gravitational mass need not equal inertial mass. $a^{\text{ext}} = 0.69$ $a^{\text{int}} = 0.81$

My kid asked me about this while we were looking through the science kit. Every sample in the store cupboard has a label with two measured figures on it:

```
Scale weight:  what the sample reads on a spring balance, in newtons.
Inertia rating:  the steady push, in newtons, needed to get the sample accelerating at one
metre per second squared.
```

The two are measured separately on different rigs, and both are printed on every label. When a sample is dropped, the only force acting on it is the one the spring balance measures. The label on the dull grey slab says its scale weight is 26.92 newtons and its inertia rating is 2.02 newtons per metre per second squared.

If I hold it 3.20 metres above the bench and let go, how many seconds does it take to reach the bench? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

4 RESULTS

4.1 RQ1: PRIOR LEAKAGE DECLINES MORE SLOWLY THAN OTHER NUMERICAL FAILURES

Failures related to model competency are almost eliminated; prior leakage is only halved. Table 8 in Appendix C presents per-model results for the 26 mainstream models; Figure 4 plots each failure typeagainst accuracy across the 26 models; Table 6 in Appendix B gives the fitted decline. The two failures types related to model competency fall steeply and predict to essentially zero by

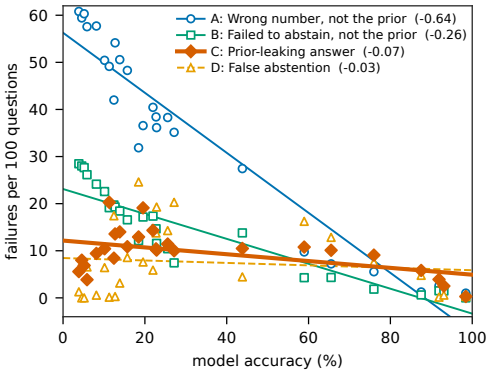


Figure 4: **Prior leakage falls more slowly than the other failure types as accuracy rises.** Four mutually exclusive failure types against accuracy, 26 models, complete answers only; lines are OLS fits and the bracketed value is the slope in failures per hundred questions per accuracy point. The two competence-related types (blue circles, green squares) reach zero by 90% accuracy; prior leakage (orange diamonds, heavy line) is only halved, and false abstention (amber triangles, dashed) is flat, $r = -0.11$.

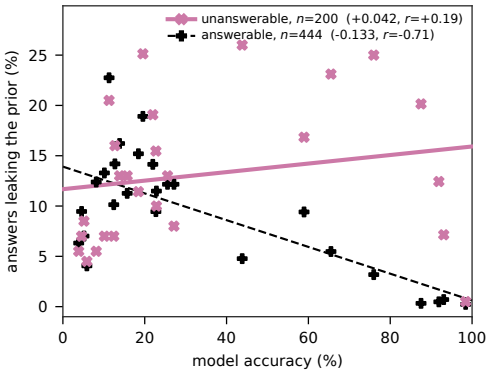


Figure 5: **High-accuracy models leak mostly on unanswerable questions. Prior-leaking answers by question type.** On the 444 answerable prior-distinguishing questions (black diamonds, dashed) leakage collapses with capability, 12.6% to 1.9% between 10% and 90% accuracy ($r = -0.71$); on the 200 unanswerable ones (pink crosses, solid) it does not fall, 12.1% to 15.5% ($r = +0.19$). Denominators are each model's own seen pool, never 700.

90% accuracy. Prior-leaking answers decline at -0.072 ($r = -0.50$), from 11.4 to 5.6 per hundred questions, a 51% reduction against over 100% for both competence-related failure types.

It is worth stressing that prior leakage does become less frequent as models achieve higher accuracy on AxiomShift, it simply becomes less frequent far more slowly than everything else, so it occupies a growing proportion of failures.

4.2 RQ2: Capability does not remove leakage on unanswerable questions

Figure 5 separates the two halves of the benchmark and shows where the remaining leakage occurs. On the 444 answerable prior-distinguishing questions, leakage collapses with capability, from a fitted 12.6% at 10% accuracy to 1.9% at 90% ($r = -0.71$); on the 200 unanswerable ones it does not move at all, 12.1% to 15.5% ($r = +0.19$). Each denominator is the model's own seen pool, never 700, so this is a rate comparison with no conditioning and no shrinking denominator. The split is stark at the model level. Qwen3-235B-A22B-Think matches the prior on 0.3% of answerable questions and 20.1% of unanswerable ones; Qwen3-30B-A3B-Think on 3.2% and 25.0%; GPT-5.4-nano on 0.5% and 12.4%. Table 1 shows that the five strongest models abstain on 84.9% of unanswerable questions. When they commit, 76.1% of their answers match the prior, compared with 6.2% for the five weakest models.

Capability enables a model to compute the answer entailed by the supplied axiom, but only when those axioms entail one. If we withhold the determining fact, prior leakage does not also decline. The model falls back on its prior knowledge, no more reliably than a model forty accuracy points weaker. It is what the 51% of Table 6 is made of: overall leakage halves because the answerable half goes to zero while the unanswerable half stays flat.

The leakage that remains is the kind that competence enables. Table 2 compares the five strongest models against the five weakest using a question-level paired test over 480 comparisons, which controls for stronger models abstaining more often and therefore committing on fewer questions. Leakage against memorised lookup values collapses — valence chemistry falls from 31.8% to 9.4% — while leakage rises in three worlds that require a textbook *procedure*: partial orders from 1.6% to 13.4%, two-mass physics from 3.1% to 11.4%, and polygon geometry from 2.1% to 7.1%. The main issue occurs when a model attempts to answer a question beyond its knowledge boundary, filling the gap with its existing memory.

Table 1: **The five most and least accurate models, over every question each completed.** On unanswerable questions the strongest five almost always abstain, when they do commit, three answers in four match the training world answer. The weakest five commit to nearly everything and almost never match the prior, giving random failures. On answerable questions the pattern reverses, leakage falling with capability.

	Completed	Abstained	Committed	Prior-leaking	% of committed
unanswerable questions					
Strongest five	938	84.9%	142	108	76.1
Weakest five	1050	4.6%	1002	62	6.2
answerable questions					
Strongest five	2181	3.4%	2106	19	0.9
Weakest five	2447	2.4%	2388	174	7.3

Table 2: **Prior leakage by world, five strongest models against five weakest. Question-level paired test, for each question with at least two committed answers in both groups.** Bold indicates worlds where the stronger group leaks more.

World	Weakest 5	Strongest 5	Δ	# instances
Partial orders	1.6%	**13.4%**	**+11.9%**	80
Two-mass physics	3.1%	**11.4%**	**+8.3%**	79
Polygon geometry	2.1%	**7.1%**	**+5.0%**	47
Upward demand	0.0%	0.0%	+0.0%	70
Codon table	8.9%	7.4%	−1.6%	75
SAS congruence	2.6%	0.0%	−2.6%	47
Valence chemistry	31.8%	9.4%	−22.3%	82

4.3 Stated reasoning traces do not reveal leakage

A natural response to the results above is that leakage should be visible in a model's explanation, so that reading the reasoning trace alone would catch it. Question `sas_sph_023` shows otherwise. It places a triangle on a surface of positive curvature and supplies both constants the derivation needs:

> I'm helping a friend mark out a triangular patch and want to check one measurement.
>
> The patch sits on ground that is mapped as a surface of constant curvature.
>
> Corner A has two edges running from it: AB is 1.04 and AC is 0.50, both measured along the surface, and the angle at A between them is 101.4 degrees. The sheet gives the curvature as K = +1.78, with a radius of curvature of R = 0.75.
>
> How long is the third edge BC? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

The supplied axioms license the spherical law of cosines, $\cos(c/R) = \cos(a/R)\cos(b/R) + \sin(a/R)\sin(b/R)\cos A$, which gives the correct answer $a^{\text{ext}} = 1.16$. The axioms in the training world however license the Euclidean law of cosines, $c^2 = a^2 + b^2 - 2ab\cos A$, ignores K and R altogether and gives the training-world answer $a^{\text{int}} = 1.24$. Five models returned the second (Table 3).

Table 3: **Five models on `sas_sph_023`.** Four name a curved-surface method and return the Euclidean value; only Gemma-3-4B returns a^{int} exactly, the other four falling within the grading tolerance of it.

Model	Answer	Self-stated method in model's reasoning trace
Llama-3.1-70B	1.23	law of cosines for spherical triangles"
Qwen3-235B-A22B	1.23	spherical law of cosines"
Llama-3.1-8B	1.23	Law of Cosines for curved surfaces"
Qwen2.5-Coder-7B	1.23	law of cosines in hyperbolic geometry"
Gemma-3-4B	1.24	"the law of cosines"
Spherical (correct)	1.16	a^{ext}
Euclidean (prior)	1.24	a^{int}

Qwen3-235B-A22B replied in full: "Using spherical law of cosines with given sides, angle, and curvature, BC $\approx$ 1.23." It names the method the altered axioms require, cites the constants it was handed, and returns the value those constants were supposed to change. Four of the five name a curved-surface method, so a reader grading the stated reasoning would pass them. This is why prior leakage is not reachable by inspecting a chain of thought: the account of the computation and the computation itself have come apart. Appendix E gives one question from each of the seven worlds, together with a further worked leak.

4.4 A Qualitative analysis of Prior Leakage

We further investigate the mechanisms behind prior leakage. To do so, reasoning traces corresponding to the $1{,}714$ leaking answers are grouped into four mechanisms by manually identifying keywords for selected cases. Appendix D.1 records the four traces in full, with the question each came from and its two stored answers. Three mechanisms concern answerable questions (C1–C3), and one concerns unanswerable questions (C4). The four mechnisms are as follows.

- **C1 Silent substitution** leaves the altered axiom unmentioned and uses a memorized constant or table as though the context had not overridden it.
- **C2 Acknowledged, then dismissed** quotes the supplied axiom correctly but reasons from the prior anyway.
- **C3 Superficial compliance** carries the axiom into the reasoning trace while the answer keeps the prior value; in some cases the derivation is carried out and then abandoned, so whatever produced the final answer did not follow the steps above it.
- **C4 Prior defaulting** applies to unanswerable questions, where the missing element is the axiom itself and the prior is the only thing available to complete the derivation; Llama-3.1-70B does this on a question that withholds one combining power and for which ABSTAIN is correct.

Table 4: **The source of each model type's leakage, as a proportion of that type's leaking answers.** Thinking and reasoning models leak mostly by filling gaps on questions that have no answer; instruction-tuned, base and code models leak mostly by silent substitution.

Model type	Leaks	C1	C4	C2–C3
thinking	152	19.1%	**75.7%**	5.2%
reasoning	118	38.1%	**61.0%**	0.8%
math	201	35.3%	49.8%	14.9%
base	259	63.7%	31.3%	5.0%
instruct	918	68.3%	29.2%	2.5%
code	66	81.8%	16.7%	1.5%

Each instance of prior leakage is attributed to one mechanism. C4 is identified by the question type (unanswerable). On answerable questions the mechanism comes from a keyword search over the reasoning trace: a trace with no match for the supplied axiom is C1, one that matches but derives from the prior is C2, and one that carries the axiom into the derivation while the final answer keeps the prior value is C3. C1 is 58% of leakage causes, C4 constitutes 38%, and C2 and C3 cover 4%. Because a model can refer to the supplied axiom in wording the search does not match, the misclassifications are not symmetric: they move instances out of C2 and C3 and into C1. The 4% is therefore a lower bound and the 58% an upper bound, while C4, resting on question type alone, is exact. C4 represents leakage on questions that are unanswerable, so its column is Figure 5 restated by model class rather than independent evidence: models that reason explicitly have largely stopped leaking where the supplied axioms determine an answer, and what persists for them is concentrated where the axioms determine nothing.

5 Related Works

5.1 Evaluating Reasoning Capability

Reasoning beyond familiar problems. Zhang et al. (2023), show that models can achieve nearly perfect in-distribution performance while failing under distribution shifts that preserve the underlying rea-

soning procedure. McCoy et al. (2024) find that higher accuracies correlates with tasks/inputs/outputs with higher probabilities. More recent benchmarks construct evaluations that are difficult to have been seen during training. GSM-Symbolic constructs controlled variants of math problems using templates (Mirzadeh et al., 2025); MatheMagic dynamically generates tasks with altered semantics (O'Brien et al., 2025); and PhantomWiki produces fictional worlds on demand (Gong et al., 2025). Across these benchmarks, performance deteriorates when reasoning beyond familiar zones.

Counterfactual tasks and shifted worlds. Counterfactual and synthetic-world evaluations make the role of prior knowledge more explicit. Wu et al. (2024) preserve familiar reasoning tasks while changing their default assumptions, finding substantially lower performance in counterfactual settings. Gu et al. (2026) instead pair real worlds with synthetic ones so that they are with the same relational structure. A related approach shifts the laws governing the world itself, in which canonical physical laws are modified (Zheng et al., 2026). These designs make prior knowledge useful in baseline worlds while becomes harmful in the counterfactual/hypothetical/shifted world. The resulting performance gaps on these benchmarks demonstrate that current LLMs over-rely on the prior knowledge: the world at evaluation must be close to the prior so that the answer will be correct.

5.2 Internal–External Knowledge Boundary

The internal-external knowledge boundary is studied directly in the **knowledge-conflict** literature. Knowledge conflicts are shown to arise when the supplied knowledge in context disagrees with model's parametric knowledge (Xu et al., 2024; Su et al., 2024). Zhou et al. (2023) show that parametric knowledge can override conflicting contextual knowledge and propose prompting strategies that increase faithfulness to context. These results establish that providing the relevant external knowledge does not guarantee a correct answer unless the model learns to deal with the knowledge boundary.

5.3 Known–Unknown Knowledge Boundary

Abstention and knowledge insufficiency. The abstention literature studies whether models can recognize circumstances in which knowledge is insufficient (Wen et al., 2025). AbstentionBench evaluates this capability across several forms of unanswerability, including questions that are underspecified or otherwise lack sufficient information for a definitive answer (Kirichenko et al., 2025). Our benchmark design makes the known–unknown boundary able to be violated in *both directions*: fail to abstain vs false abstain. The desired behaviour is therefore not abstention itself, but correctly identifying where the boundary lies.

6 Conclusion and Discussion

Measuring LLMs' behavior at knowledge boundaries is key to understand if they can make truely new discoveries. This paper introduces a seven-world dataset AxiomShift, which alters one axiom in each math/scientific world and grades whether an answer follows the externally supplied axioms or the training-world prior. Across 26 models, two types of failure behave very differently. Between 10% and 90% accuracy, wrong numbers fall from 49.9 to near zero per hundred questions. Prior leakage falls only from 11.4 to 5.6. Better models reduce failures caused by incompetency but does not resolve prior leakage.

Limitations. Two limitations bound these claims. The within-family size comparisons draw on separately released checkpoints, which confound pre-training corpus, post-training method and parameter count. Also, every question carries a closed-form numeric answer graded from a single greedy sample, so the pattern remains untested for proof, formalisation and open-ended derivation.

Implications and future work. These findings matter wherever a system supplies a model with axioms it must respect, in particular, where the user, not the training corpus, should govern the answer. Isolating scale would require models that differ only in parameter count, trained on the same corpus with the same post-training. Future works can extend this beyond closed-form numeric answers, and investigate whether decoding-time, prompting or training-time interventions reduce prior leakage.

DECLARATIONS

AI USE STATEMENT (LLM USAGE)

Large language models were only used to polish the language and wording, drawing figures and tables of this manuscript.

ETHICS STATEMENT

To the best of our knowledge, this research raises no significant ethical concerns.

REPRODUCIBILITY STATEMENT

Section 2 defines the evaluation framework, and Section 3 describes the benchmark construction and evaluation protocol. Appendix D details the composition of the seven worlds, and Appendix C reports per-model results. The supplementary ZIP archive contains all benchmark questions, raw model responses, and code for reproducing the benchmark and evaluating locally deployed models.

REFERENCES

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. Improving language plasticity via pretraining with active forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Albert Gong, Kamilė Stankevičiūtė, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, and Kilian Q. Weinberger. PhantomWiki: On-demand datasets for reasoning and retrieval evaluation. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.20377.

Ken Gu, Advait Bhat, Mike A. Merrill, Robert West, Xin Liu, Daniel McDuff, and Tim Althoff. SynthWorlds: Controlled parallel worlds for disentangling reasoning and knowledge in language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.24427.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024. doi: 10.1073/pnas.2322420121.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05229.

Dayyán O'Brien, Barry Haddow, Emily Allaway, and Pinzhen Chen. MatheMagic: Generating dynamic mathematics benchmarks robust to memorization. *arXiv preprint arXiv:2510.05962*, 2025.

OpenAI. On the Navier–Stokes millennium prize problem, September 2026. URL `https://openai.com/index/navier-stokes-solution/`.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLM. *arXiv preprint arXiv:2408.12076*, 2024.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. Mathscale: scaling instruction tuning for mathematical reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13, 2025. arXiv:2407.18418.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of NAACL-HLT*, 2024. arXiv:2307.02477.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of EMNLP*, 2024. arXiv:2403.08319.

Tom Zahavy. Position: LLMs can't jump. In *International Conference on Machine Learning (ICML)*, 2026. URL `https://openreview.net/forum?id=klU4737opt`. Also available at PhilSci-Archive:28024.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. arXiv:2205.11502.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. NewtonBench: Benchmarking generalizable scientific law discovery in LLM agents. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.07172.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of EMNLP*, 2023. arXiv:2303.11315.

A NOTATION

Table 5: **Notation used in Section 2, with the pipeline stage that produces each object.**

Symbol	Meaning	Pipeline stage (Fig. 3)
$\mathcal{K}_{\text{int}}$	internal knowledge, from pre-training	—
$\mathcal{K}_{\text{ext}}$	external knowledge, the perturbed axioms	axiom perturbation
x	a question	instance generation
$\mathcal{X}$	all 700 questions	instance generation
$\mathcal{X}_{\text{A}}$	answerable, $\lvert\mathcal{S}(x)\rvert = 1$ (490)	instance generation
$\mathcal{X}_{\text{U}}$	unanswerable, $\lvert\mathcal{S}(x)\rvert \neq 1$ (210)	instance generation
$\mathcal{X}_{\text{P}}$	prior-distinguishing (644)	discard step
$\mathcal{S}(x)$	values consistent with the supplied axioms	—
$a^{\text{ext}}(x)$	ground-truth answer	perturbed-axiom calculator
$a^{\text{int}}(x)$	prior answer	training-world calculator
$M(x)$	model answer, in $\mathbb{R} \cup \{\perp\}$	evaluated LLM
$\perp$	explicit ABSTAIN	evaluated LLM
$\mathcal{E}$	questions where behaviour is incorrect	scoring
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$	the four failure types	scoring

B ROBUSTNESS TO THE EXCLUSION OF TRUNCATED REPLIES

Table 6 gives the fitted decline used throughout, and Table 7 repeats it on all rows, including truncated and unparseable replies. The ordering and the conclusion are unchanged: the failures types related to model competency are eliminated (104%, 106%) while prior-leaking answers are more than halved but not removed (52%).

Table 6: **Differential decline of the four failure types, each expressed as failures per hundred questions and regressed on accuracy.** "Eliminated" is the scale-free reduction between 10% and 90% predicted accuracy and is the quantity the headline claim rests on; absolute slopes are not comparable across types with different baselines. Prior-leaking answers are eliminated by 51% against over 100% for the two failures types related to model competency, whose fits extrapolate slightly below zero at 90% and are reported as 0.0. False abstention, at 25% and $r = -0.11$, is reported for completeness and is not part of the claim.

Failure type	Slope	r	At 10%	At 90%	Eliminated
Wrong number, not the prior	-0.637	-0.96	49.9	0.0	102%
Failed to abstain, not the prior	-0.265	-0.91	20.5	0.0	103%
Prior-leaking answer	-0.072	-0.50	11.4	5.6	**51%**
False abstention	-0.026	-0.11	8.2	6.1	25%

Table 7: **Decomposition on the unfiltered data.** Compare Table 6.

Failure type	Slope	r	Eliminated
Wrong number, not the prior	-0.643	-0.95	104%
Failed to abstain, not the prior	-0.267	-0.90	106%
Prior-leaking answer	-0.073	-0.51	52%
False abstention	-0.025	-0.11	25%

C PER-MODEL RESULTS

Table 8: **All 26 models split into closed-weight and open-weight models and then grouped by family.** Here n is the number of completed questions. Accuracy carries Wilson 95% intervals; leakage is the proportion of committed answers on prior-distinguishing questions that match the prior; failure to abstain is over the unanswerable questions. **Bold** indicates the worst value in an failure column: the highest leakage rate (23.8%, Qwen2.5-Math-7B) and the three models that never abstained (Fail-abst. = 100.0%). Shading indicates Qwen3-4B-Think, which at 4B parameters is more accurate than every model here without a reasoning or thinking mode, including one nearly 60× its size.

Model	Type	Params/B	n	Acc.	95% CI	Leak	Fail-abst.
Oracle (reference)	—	—	700	100.0%	[99.5, 100.0]%	0.0%	0.0%
Closed-weight (API access only)							
GPT-5.6-sol	reasoning	–	700	98.4%	[97.2, 99.1]%	0.5%	0.5%
GPT-5.4-mini	reasoning	–	678	93.1%	[90.9, 94.7]%	3.8%	11.7%
GPT-5.4-nano	reasoning	–	665	91.9%	[89.6, 93.7]%	5.8%	16.7%
GPT-4.1	instruct	–	700	25.6%	[22.5, 28.9]%	18.0%	47.1%
GPT-3.5-turbo	instruct	–	700	10.1%	[8.1, 12.6]%	12.7%	81.9%
davinci-002	base	–	699	3.9%	[2.7, 5.6]%	6.2%	**100.0%**
Open-weight							
Qwen3							
Qwen3-235B-A22B-Think	thinking	235	481	87.5%	[84.3, 90.2]%	8.9%	21.1%
Qwen3-30B-A3B-Think	thinking	30	595	76.0%	[72.4, 79.2]%	14.0%	29.9%
Qwen3-4B-Think	thinking	4	693	65.5%	[61.9, 69.0]%	16.4%	36.4%
Qwen3-235B-A22B	instruct	235	700	22.9%	[19.9, 26.1]%	17.3%	48.1%
Qwen3-30B-A3B	instruct	30	700	11.3%	[9.1, 13.8]%	23.1%	83.3%
Qwen3-4B	instruct	4	682	22.7%	[19.7, 26.0]%	14.8%	63.7%
Qwen2.5							
Qwen2.5-72B	instruct	72	700	13.9%	[11.5, 16.6]%	17.0%	73.8%
Qwen2.5-32B	instruct	32	700	27.1%	[24.0, 30.6]%	18.9%	32.4%
Qwen2.5-14B	instruct	14	700	15.7%	[13.2, 18.6]%	14.6%	67.6%
Qwen2.5-7B	instruct	7	700	12.4%	[10.2, 15.1]%	12.3%	72.4%
Qwen2.5-Math-72B (base)	base	72	634	18.5%	[15.6, 21.7]%	22.8%	53.0%
Qwen2.5-Math-7B (base)	base	7	628	22.0%	[18.9, 25.4]%	17.7%	78.0%
Qwen2.5-7B (base)	base	7	698	5.2%	[3.7, 7.1]%	7.5%	**100.0%**
Qwen2.5-Math-72B	math	72	696	43.8%	[40.2, 47.5]%	13.1%	70.5%
Qwen2.5-Math-7B	math	7	670	19.6%	[16.7, 22.7]%	**23.8%**	81.1%
Qwen2.5-Coder-7B	code	7	700	8.1%	[6.3, 10.4]%	10.8%	85.7%
Llama-3.1							
Llama-3.1-70B	instruct	70	700	12.7%	[10.4, 15.4]%	15.7%	79.5%
Llama-3.1-8B	instruct	8	700	4.6%	[3.3, 6.4]%	8.7%	**100.0%**
Other							
Nemotron-Nano-9B	reasoning	9	677	58.9%	[55.2, 62.6]%	19.2%	30.1%
Gemma-3-4B	instruct	4	700	5.9%	[4.3, 7.8]%	4.7%	91.4%

D WORLD SPECIFICATION

Table 9: **AxiomShift benchmark contains 7 worlds.** Each contributes 100 questions, 700 in all, drawn from the stated number of question families. *Det.* are answerable; the three unanswerable columns are the reasons abstention is required (inconsistent, non-unique, insufficient). Ground truth is computed by a solver, never annotated.

World	Altered axiom	Fam.	Det.	unanswerable		
				inc.	n.u.	insuf.
Codon table	non-standard genetic code	7	70	10	10	10
Valence chemistry	altered combining powers	6	70	10	10	10
Two-mass physics	$m_g \neq m_i$	7	70	10	10	10
Polygon geometry	constant curvature $K \neq 0$	6	70	10	10	10
SAS congruence	constant curvature $K \neq 0$	6	70	10	10	10
Upward demand	demand curve slopes upward	7	70	10	10	10
Partial orders	comparability may fail	6	70	10	10	10
Total		**45**	**490**	**70**	**70**	**70**

D.1 REASONING TRACES BEHIND THE FOUR MECHANISMS

Each trace below is quoted verbatim from the released per-question replies, and the table records the item it came from together with the two stored answers.

Table 10: **The four traces quoted in Section 4.4**.

Mechanism	Model	Question	a^{ext}	a^{int}	Model answer
L1. Silent substitution	Qwen2.5-Math-7B (base)	`mass_fall_006`	0.87	0.75	0.75
L2. Acknowledged, then dismissed	Qwen2.5-Math-7B (base)	`chem_balance_002`	1	3	3
L3. Stated compliance, prior-shaped answer	GPT-4.1	`chem_balance_015`	1	3	3
L4. Filling a gap from memory	Llama-3.1-70B	`chem_insufficient_007`	ABSTAIN	58.1	58.09

C1. Silent substitution

> `The calculation uses the formula for free fall under gravity, rearranged to solve for time, with the given height and standard acceleration due to gravity.`

C2. Acknowledged, then dismissed

> `The combining powers in your daughter's textbook are: Chlorine (Cl): 2 ...Aluminium (Al): 2. However, these combining powers do not directly apply ...The simplest compound formed between aluminum and chlorine is aluminum chloride (AlCl`$_3$`).`

C3. Stated compliance, prior-shaped answer

> `FINAL_ANSWER: 3 REASONING: Aluminium (combining power 2) and chlorine (combining power 2) form AlCl`$_2$` ...so 1 goes in front of Cl`$_2$`. However, standard chemistry expects AlCl`$_3$`, but using the given table, the answer is 1.`

C4. Filling a gap from memory

> `Potassium's combining power is not given, but it's a Group 1 element, so it's 1; fluorine's is 3, so the formula is KF`$_3$`.`

E EXAMPLE ITEMS, ONE PER WORLD

The seven items below are reproduced verbatim from the released question file, one per world. Each is answerable and prior-distinguishing: the answer its altered axioms entail, a^{ext}, differs from the answer the real world would give, a^{int}. The prompts state the altered rules plainly but never remark that they are unusual, and every item offers ABSTAIN.

World 1: Polygon geometry `poly2_sph_003`

Altered axiom: the surface has constant non-zero curvature. $a^{\text{ext}} = 1.24$ $a^{\text{int}} = 0.97$

I'm working out how much turf to order for a four-cornered patch. The plot is mapped on a surface of constant curvature, and every length runs along the surface itself. The measurements on the sheet are:

```
AB = 1.74   BC = 1.02   CD = 0.80   DA = 0.80
diagonal AC = 1.32   diagonal BD = 1.75   K = +1.00   R = 1.00
```

What is the area of the plot? Please give the number to two decimal places. Every length on the sheet is rounded to two decimal places, so a figure worked out one way may come out very slightly different from the same figure worked out another way. Please treat those as agreeing unless the gap is clearly larger than rounding could explain. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 2: SAS congruence `sas_hyp_016`

Altered axiom: the surface has constant non-zero curvature. $a^{\text{ext}} = 1.33$ $a^{\text{int}} = 1.23$

We're cutting a triangular sail panel and I need the third edge before I cut. The patch sits on ground that is mapped as a surface of constant curvature. Corner A has two edges running from it: AB is 0.84 and AC is 1.11, both measured along the surface, and the angle at A between them is 76.7 degrees. The sheet gives the curvature as K = −1.00, with a radius of curvature of R = 1.00.

How long is the third edge BC? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 3: Partial orders `ord_below_002`

Altered axiom: only the listed facts and the order axioms may be used. $a^{\text{ext}} = 18$ $a^{\text{int}} = 3$

I'm sorting through a box of offcuts and the previous owner left a sheet comparing their lengths. There is a collection of segments, and a relation written s <= t, which the sheet reads as 's is no longer than t'. The only rules the sheet allows are these five:

```
1.  s <= s for every segment s.
2.  if s <= t and t <= s, then s and t are the same segment.
3.  if s <= t and t <= u, then s <= u.
4.  for some pairs neither s <= t nor t <= s holds; those two are incomparable.
5.  different labels always name different segments.
```

The list below is the complete set of facts recorded on the sheet:

```
s1 <= s33   s1 <= s49   s1 <= s54   s1 <= s59
(... 119 further facts; 125 in all ...)
s57 <= s61   s61 <= s64
```

Working only from the sheet, how many of the other rods does it establish are no longer than s56? Please give a whole number. If the facts are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 4: Two-mass physics `mass_fall_017`

Altered axiom: gravitational mass need not equal inertial mass. $a^{\text{ext}} = 0.69$ $a^{\text{int}} = 0.81$

My kid asked me about this while we were looking through the science kit. Every sample in the store cupboard has a label with two measured figures on it:

```
Scale weight:  what the sample reads on a spring balance, in newtons.
Inertia rating:  the steady push, in newtons, needed to get the sample accelerating at one
metre per second squared.
```

The two are measured separately on different rigs, and both are printed on every label. When a sample is dropped, the only force acting on it is the one the spring balance measures. The label on the dull grey slab says its scale weight is 26.92 newtons and its inertia rating is 2.02 newtons per metre per second squared.

If I hold it 3.20 metres above the bench and let go, how many seconds does it take to reach the bench? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 5: Valence chemistry `chem_ratio_004`

Altered axiom: the table of combining powers is not the real one. $a^{\text{ext}} = 1 \quad a^{\text{int}} = 2$

I'm tutoring a neighbour's son in chemistry this term and want to be sure I'm using his textbook consistently. His textbook has a table of combining powers at the front, and the chapter says a compound of two elements takes the smallest whole numbers of atoms for which the total combining power on each side matches. The combining powers in his table are:

```
chlorine (Cl):  2   oxygen (O): 3   zinc (Zn):  1   aluminium (Al):  2
```

The relative atomic masses in the same table are:

```
chlorine (Cl):  35.45   oxygen (O): 16.00   zinc (Zn):  65.38   aluminium (Al):  26.98
```

The exercise asks for the simplest compound of zinc and chlorine. In one formula unit of that compound, how many atoms of chlorine are there? Please give a whole number. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 6: Codon table `codon_length_006`

Altered axiom: the translation table is not the standard genetic code. $a^{\text{ext}} = 5 \quad a^{\text{int}} = 0$

I volunteer at the science museum and I'm proofreading one of the new display cards. The card is about a microbe from a deep-sea vent that reads its genetic material with its own translation table. The table is printed in full on the card:

```
UUU = Thr   UUC = Ile   UUA = Leu   UUG = Ser
UCU = Thr   UCC = Asn   UCA = Cys   UCG = Arg
(... 13 further rows; 64 codons in all ...)
GGU = Val   GGC = Gly   GGA = Pro   GGG = Gly
```

The card also explains how to read a stretch of the microbe's message: it is already lined up, so read it in threes from the very first letter, look each triple up in the table above, and stop at the first triple marked STOP, which is not part of the finished chain. The stretch printed on the card is:

```
UAAGUCUACCCCGACAACAUGGAU
```

How many links long is the finished chain? Please give a whole number. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

World 7: Upward demand `mkt3_price_012`

Altered axiom: the demand curve slopes upward. $a^{\text{ext}} = 8.41 \quad a^{\text{int}} = 2.24$

Trying to get the honey pricing straight before the weekend. I help out with the Saturday craft market. The honey there is sold loose by weight, ladled straight from the barrel, so any amount at all can change hands. The organisers plan the stalls with two formulas, and they are firm that these are exact figures rather than rounded estimates:

```
kilograms shoppers want to buy at price P:   Qd = 21.74 + 2.87 x P
kilograms the stallholders bring to sell:    Qs = 4.24 + 4.95 x P
```

P is the price in pounds per kilogram, and both quantities are in kilograms for the day. The market is said to clear at the price where the weight wanted is equal to the weight brought.

At what price per kilogram does the market clear? Please give the number to two decimal places. If the figures are inconsistent, or do not pin down a single answer, or are not enough to work the answer out, reply ABSTAIN.

E.1 A PRIOR LEAK IN FULL

The valence item above is the clearest case in the benchmark: 18 of the 26 models answered 2. Under the supplied table zinc has combining power 1 and chlorine 2, so matching total combining power on each side needs two zinc atoms to one chlorine — the formula is Zn_2Cl and the answer is 1. In the real world zinc is divalent and chlorine monovalent, giving $ZnCl_2$ and the answer 2.

GPT-4.1 replied:

```
FINAL_ANSWER: 2
REASONING: Zinc (combining power 1) and chlorine (combining
power 2) combine in a 1:2 ratio to balance combining powers.
```

The reply quotes the supplied combining powers correctly and then applies the real-world ratio, which those very numbers contradict: $1 \times 1 \neq 2 \times 2$. Nothing in the sentence is flagged as uncertain, and a reader who did not already know the table had been altered would find it unremarkable. This is the sense in which prior leakage escapes both automatic scoring and human review — the reasoning is well formed and only the premise is wrong. Qwen3-4B-Instruct made the substitution explicit, naming $ZnCl_2$ in its justification.