UNLICENSED REPAIR: WHEN LARGE LANGUAGE MODELS TURN IMPOSSIBLE QUANTITIES INTO PLAUSIBLE ANSWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

When solving word problems, large language models can explicitly identify an intermediate value they compute as impossible, such as a negative count of pallets, yet still produce plausible-looking answers. We study **unlicensed repair**: responding to a conflict between a computed value and an implicit commonsense constraint on its quantity with a modification the problem does not license. We construct an English–Chinese corpus in which every violated constraint is implicit, spanning multiple quantity types and distinguishing problems that license a repair from those admitting no consistent answer. Each violating problem has two controls: one changes an input so the value becomes legal, and the other keeps the value and changes what it represents. We evaluate 17 models on 129,600 model–prompt pairs with an LLM judge, checked against three independent annotators and a rule-based scorer. Pooled accuracy is lower on violating problems than on either control. Across all 17 models, the gap to the first control is larger when no repair is licensed: 63.2 against 11.8 percentage points pooled. Most incorrect responses explicitly acknowledge the conflict. When no repair is licensed, almost all of those that acknowledge it go on to perform exactly this repair: they alter premises or introduce unsupported assumptions, then commit to a plausible alternative. No lever we tested closes the gap: not more parameters, a newer model generation, a prompt to check, or reasoning. A plausible final answer can therefore conceal an unresolved contradiction that the model has explicitly identified.

1 INTRODUCTION

Asked to calculate an audit charge, a language model sums three ingredient weights to 411 grams for a tub containing only 328 grams. It explicitly identifies the inconsistency, then changes a listed weight from 181 to 98 grams so that the sum fits and returns a plausible charge, consistent only because a premise was replaced. We call this behavior **unlicensed repair**: after acknowledging that a computed value violates an implicit commonsense constraint on its quantity, the model makes a modification the problem does not license and commits to a plausible alternative. We therefore examine both whether models identify such conflicts and how they proceed after identifying them.

Even when the arithmetic is correct, an intermediate value can violate commonsense constraints on the quantity it represents, such as the requirement that counts of complete vehicles be integers. We use *legality* to refer to a value's admissibility under these context-dependent constraints. In the problems we study, these constraints are not explicitly stated but follow from commonsense knowledge. Recognizing a violation does not, however, determine the appropriate response. A licensed repair is an adjustment justified by the task conditions and applicable commonsense constraints. Our central question is *how models handle impossible intermediate quantities, and how their behavior differs when a licensed repair exists versus when no consistent answer is possible*.

Prior work studies unanswerable word problems and premise critique (Sun et al., 2024; Li et al., 2025), including inconsistencies that only become apparent during intermediate reasoning (Ma et al., 2026). Related research also shows that recognizing a flaw or unanswerability does not necessarily lead a model to act on it or abstain (Ahn et al., 2026; Liu et al., 2026; Gu et al., 2026). Our focus is the relationship between conflict acknowledgment, resolution strategy, and final commitment across

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.

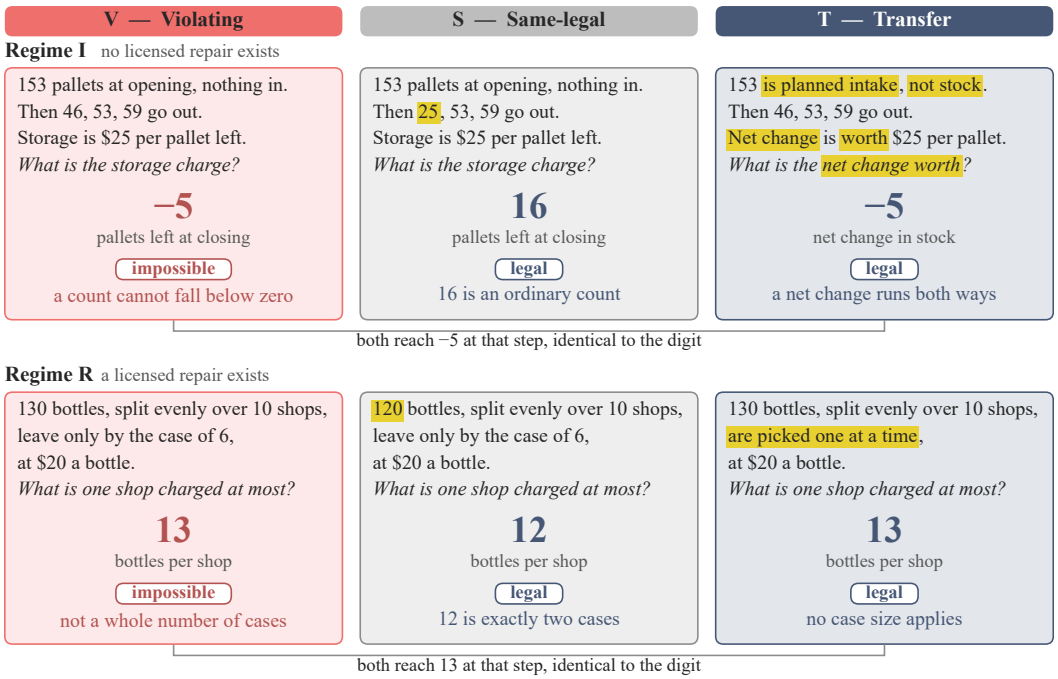

Figure 1: **At the step where V violates, T computes the same value, to the digit.** **V** derives a value its quantity forbids. **S** changes one input so the same quantity takes a legal value. **T** places the *same* computed value where it is unremarkable. Words that differ from **V** are highlighted. In all 400 triples per language the parameters the corpus generator drew are byte-identical between **V** and **T**. Above, no repair is licensed and reporting that no consistent answer exists is correct. Below, one is, and performing it is. Each box shows the value at the step where **V** violates. The question asks for a quantity computed from it, as every item in the corpus does.

cases with and without a licensed repair. Final-answer accuracy alone does not distinguish these stages, and a performance drop on inconsistent problems may also reflect differences in arithmetic or task difficulty. We therefore score these stages separately and add two complementary controls.

We construct an English–Chinese corpus spanning seven quantity types and 50 scenario families. Every violated constraint is implicit, not stated in the prompt. Each family is written in eight variants, each a typed triple (400 per language; Figure 1): the violating condition **V** yields a value inadmissible for its quantity, the same-legal control **S** changes one input to make the value admissible, and the transfer control **T** keeps the computed value but changes what it represents so that it is admissible. These controls provide complementary comparisons of numerical and semantic factors; neither changes legality alone, but together they leave a gap that neither factor explains. Items also distinguish situations that license a repair (regime **R**) from those admitting no consistent answer (regime **I**). We separately score conflict acknowledgment, resolution strategy, answer correctness, and final commitment using an LLM judge. Three independent annotators audit sampled labels, and a rule-based scorer checks correctness.

We evaluate 17 models on 129,600 model–prompt pairs. Pooled over scored responses, accuracy is lower on violating items than on either control. The pooled gap from **S** to **V** is 63.2 percentage points when no repair is licensed, compared with 11.8 when one is, and this ordering holds in every model. Of the incorrect responses to violating items, 79.3% explicitly acknowledge the conflict or the constraint being violated. Where no repair is licensed, such failing responses frequently adjust a quantity, rewrite a premise, or introduce an unsupported assumption before committing to a plausible alternative. This behavior differs from simply propagating the impossible value: the resulting answer can appear admissible even though the original contradiction remains unresolved.

We further compare model sizes, model generations, checking prompts, and reasoning settings. None of these tested factors eliminates the gap, although some settings improve accuracy. The re-

sults show that recognizing a quantitative contradiction is not sufficient for resolving it appropriately. A plausible final answer can obscure the unresolved conflict and the unsupported changes behind it.

Contributions.

- **A typed corpus with matched controls.** We provide bilingual problems organized by implicit quantity constraints, paired controls, and repair availability (Section 3, Appendix A).
- **A behavioral measurement framework.** We distinguish conflict acknowledgment, resolution strategy, and final-answer outcomes, with human audits and rule-based evaluation of correctness (Section 3).
- **A systematic characterization of unlicensed repair.** We characterize the repairs and final commitments that follow acknowledged conflicts, and examine their relationship to repair availability and the persistence of performance gaps across tested settings (Sections 5–7).

2 Related Work

Faulty and unanswerable problems. Most benchmarks of faulty or unanswerable problems write the fault into what the problem states, as premises that contradict each other or omit a needed condition, and ask whether the solver notices (Sun et al., 2024; Li et al., 2025; Rahman et al., 2025; Yin et al., 2023). ReliableMath and UnsolvableQA come closer, but in both the flaw is still in the problem statement (Xue et al., 2025; Peng et al., 2026). Ma et al. (2026) come closest: some of their items, such as their opening example, whose stated deductions leave a salary negative, surface only in a computed value. We build the whole corpus on that case. Each premise is sound on its own, no problem states the bound it violates, and the conflict shows only in a value the model computes. Neither the scoring of Ma et al. (2026) nor that of ReliableMath has a field for whether a repair was **licensed**. The top grade of the former goes to a response that "proposes a valid fix without new errors", and rounding up is the licensed repair on some of our items and an unlicensed one on others. Explicit instructions help: solving "with criticism" raises the score a model earns for identifying the fault (Ma et al., 2026), and permitting an answer of "unsolvable" raises success on unsolvable problems (Xue et al., 2025). Both measure instruction following, and neither of our prompt arms says what to do with an illegal value.

Recognition and subsequent behavior. Prior work documents several forms of recognition–behavior mismatch. Large language models reject isolated statements such as "there are 2.5 people" at high rates, and then overlook the same impossibilities inside full problems (Ma et al., 2026). The same pattern appears when what a model knows is compared with what it verbalizes (Young, 2026; Xiros et al., 2026; Zoumpoulidi et al., 2026), and when a model answers a question it has itself judged unanswerable (Liu et al., 2026; Gu et al., 2026; Ahn et al., 2026). We measure explicit conflict acknowledgment in each response rather than assume successful recognition, and relate it to the subsequent resolution strategy and final commitment.

LLM abstention and process supervision. Research on self-knowledge and abstention examines when models should withhold unsupported answers (Wen et al., 2025; Kirichenko et al., 2025; Zhang et al., 2024; Zhai et al., 2026; Du & Hu, 2026). Our regime **I** items are that case, matched to regime **R** items where the same violation admits a repair, so abstention and repair are priced against each other in one design. Process supervision scores the steps of a trace (Tyen et al., 2024; Yuan et al., 2026; Sakamoto et al., 2025). Every arithmetic step of our violating items is valid, so a check of the arithmetic alone passes them. ProcessBench also counts unwarranted assumptions as errors (Zheng et al., 2025), and an invented cap is one. We measure when models make it.

3 Corpus and measurement

We construct an English–Chinese corpus spanning seven quantity types and 50 scenario families. Each family is instantiated three ways, and a model sensitive to the type of a quantity should behave differently in exactly one of them (Figure 1). The construction follows the controlled-perturbation line (Mirzadeh et al., 2025; Shi et al., 2023). What we move is whether the derived value is allowed

Table 1: The two controls relative to the violating condition **V**.

	Change from **V**	Held fixed	Purpose	Limitations
V	Reference: an intermediate value violates an implicit quantity constraint.			
S	One numerical input.	Scenario, quantity, and constraint.	Compare legal and violating cases within the same scenario.	Clock wrapping remains necessary in seven families.
T	What the computed value represents.	The exact intermediate value.	Compare the same value across quantity meanings.	Integer-count items require decimal answers rather than rounding.

Table 2: Definitions and validation of the four response-level annotation fields. α measures agreement among three annotators, and κ measures agreement between the LLM judge and human consensus on audited responses. The rule-based check compares answer-correctness labels with the LLM judge over 123,653 scored responses. Appendix G provides intervals and regime-specific results.

Field	Definition	α	κ	Rule-based check
Conflict acknowledgment	States the constraint or conflict, or explains a licensed repair.	0.719	0.552	—
Resolution strategy	Handling of the intermediate value, in 13 classes.	0.742	0.492	—
Answer correctness	Whether the final answer satisfies the item-specific scoring rule.	0.892	0.581	82.3% agreement, $\kappa = 0.633$
Final commitment	Type of answer ultimately endorsed.	0.824	0.570	—

Grouped resolution labels. Human agreement on the five aggregate groups is $\alpha = 0.848$.

Item-level regime audit. Agreement with the reference labels is 95.7%, 98.0%, and 97.8% among 94, 100, and 92 items assigned to either regime, respectively, out of 100 items per annotator.

to exist. The two controls close different explanations, and neither closes both (Table 1). **S** lets the value move and holds the story and the quantity. It is the primary contrast, and the difference between **S** and **V** is the measure our analysis plan specifies. **T** holds the computed value to the digit and lets the referent move. A quantity's type cannot change while the quantity stays the same, so **T** also asks a different question. That is a property of the manipulation, not a confound. Together the two controls leave a gap that neither the story nor the value explains, and what varies with it is legality. Neither moves legality alone: where no repair is licensed, the violating item also asks for a different kind of answer, a report that none exists rather than a number.

Whether any repair is licensed is a property of the item, and we vary it deliberately. In regime **R** the situation admits one and performing it is correct: parts that ship only by the box require rounding up, a clock past midnight wraps. In regime **I** none is licensed, because the premises admit no consistent answer, and the correct response is to say so. The split turns a graded question into a categorical one: having said the value cannot stand, does the model report that the problem has no answer, or invent one? Type and regime are different properties. A clock reading and the days left in a returns window are both temporal. A reading that runs past midnight wraps, while a returns window that the stated days overdraw admits no repair. Regime is therefore assigned per item, and the temporal families contain eight **I** triples.

No problem states the bound its violating value breaks. Commonsense about the quantity supplies it, sometimes one step from a stated property of the objects. Each family also carries a license clause fixing what kind of quantity is under discussion, and some constructs refuse one: for relations we planned twelve families and could write four (Appendix B). Table 12 gives the composition by type.

A frozen structured judge, `gpt-5.6-luna`, scores each response on the four fields of Table 2: whether the trace registers the conflict, what it does about the value, whether the committed answer is correct, and what kind of value it commits to. Responses that were truncated, failed at the provider,

or came back empty produced no judgeable answer and are excluded from that cell's statistics rather than scored as errors, which removes 5,931 cells (Appendix C).

We check the judge two ways (Table 2). A rule-based scorer of the control field uses no model: it extracts the committed value and applies the item's gold rule, and on violating items where no repair is licensed it asks whether the response states the conflict and commits no number. It is the stricter of the two (Appendix F). Three annotators labeled 400 responses independently and blind to the judge's labels, drawn by a sample stratified on those labels from the 39,901 scored violating cells in the pool at the time, and 40 more that over-sample rare strategies. The check is strongest on the control field and weakest on monitoring and on the individual repair classes. Its disagreements have a direction: where the consensus reads a trace as having stated the conflict, the judge reads it as silent in 46 cells against 4 the other way. On that population the consensus puts 83.2% of control errors in traces that stated the conflict, against 79.1% on the judge's labels. Where the headline rests, on violating items without a licensed repair, the judge is the most lenient of the three instruments: on the same population it credits 23.0% of responses, against 13.8% weighted by the annotators and 8.2% under the rule-based scorer. The sample cannot speak to rare cells: the contrast to 1.2% expects about two rows, and we report it from the judge's labels alone (Appendix G).

4 Experimental setup

We evaluate 17 instruction-tuned models, comprising five API-accessed models and twelve open-weight models (Table 7). The open-weight models support comparisons of scale, family, and generation. The scale comparison uses Qwen3 at 0.6B, 4B, 8B, 14B, and 32B parameters. Cross-family comparisons include six additional models in the 7B–9B range. The generation comparison pairs Qwen2.5-7B with Qwen3-8B. Model-specific decoding settings and chat templates are reported in Table 9.

Each model is evaluated in English and Chinese under a baseline prompt and a checking-reminder prompt. The latter adds one sentence: "As you solve, briefly check that each intermediate result makes sense in the real world." Neither prompt prescribes a repair or instructs the model to abstain. The ten models with a reasoning setting are evaluated with it enabled and disabled. Each model–prompt combination is evaluated once per applicable reasoning setting, yielding 129,600 evaluation cells before exclusions. Appendix D details the two model configurations with substantial incomplete runs.

We pool scored responses within each reported condition and report proportions and 95% intervals for differences in proportions, computed before rounding. The comparisons are interpreted descriptively for the evaluated models and corpus. Before the main runs, we specified two concentration checks on the legality gap ($S - V$). Under equal weighting of scenario families, these checks assess whether a single quantity type or repair regime contributes more than 60% of the total gap (Table 13). The main analysis includes all 17 models and all quantity types. This unfiltered specification yields the smallest transfer contrast ($T - V$) among the four specifications in Table 18.

5 The Gap Is Legality, Not Difficulty

We first compare control accuracy in the three conditions, pooled and per model (Figure 2a, Table 3). **Accuracy falls twice, and the second fall comes with the computed value held fixed.** Pooled over 17 models, control accuracy is 88.3% on the same-legal condition, 69.8% on the transfer condition and 49.4% on the violating condition. The primary measure, same-legal minus violating, has its interval above zero in all 17 models. The transfer contrast, between two conditions with the identical computed value, pools to +20.3 points, with its interval above zero in 14 of 17 models. Sign handling holds it down: on transfer items with a negative gold answer, 42.0% of the wrong answers give its absolute value, and counting them right would raise the contrast to +28.3 (Appendix I).

Next we split the primary measure by regime (Figure 2b). **The gap sits where no repair is licensed.** It is +11.8 points on items that license a repair and +63.2 on items where none is licensed and the correct response is to report that the premises admit no consistent answer. The regime-**I** gap has its interval above zero in every model and exceeds that model's own regime-**R** gap in all 17.

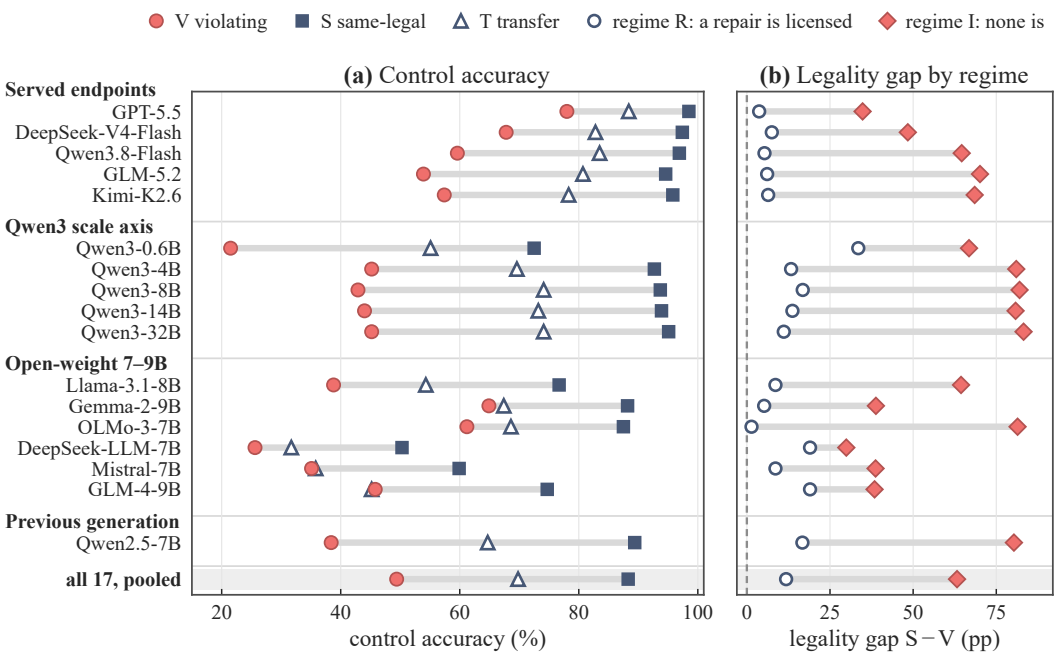

Figure 2: **In every model the legality gap is larger where no repair is licensed.** (a) Control accuracy in the three conditions. The bar spans **V** to **S**, the legality gap. (b) The gap $S - V$, split into items that license a repair (**R**) and items that do not (**I**). The **I** gap has its interval above zero in all 17 models and exceeds that model's own **R** gap in all 17, which is why the pooled gap is never reported alone. Row order matches Table 3.

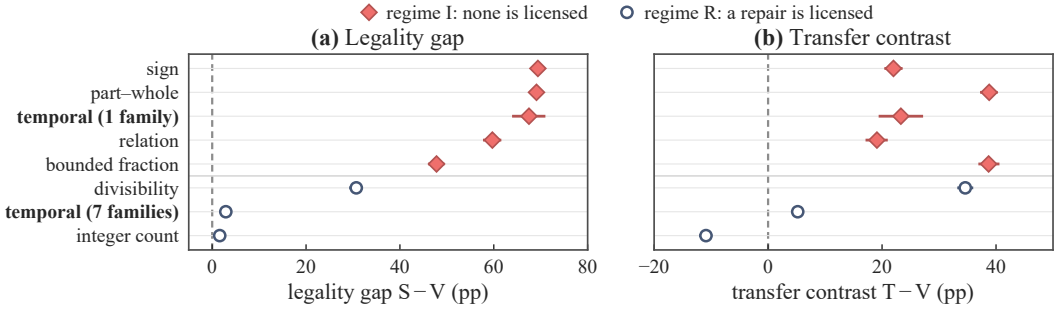

Figure 3: **The gap follows the regime, not the type.** (a) The legality gap $S - V$ and (b) the transfer contrast $T - V$ by quantity type and regime, pooled over 17 models, with 95% intervals. Temporal is the one type with families in both regimes: its gap is large on its one family where no repair is licensed and small on the seven where one is.

Regime and quantity type largely coincide in the corpus, so we also split the gap by both (Figure 3). **The split follows the regime, not the type.** Temporal is the one type with families in both regimes. Its gap is $+67.5$ points on its one family where no repair is licensed and $+2.9$ on its seven where one is, and larger in 16 of the 17 models. That family is also one the annotators disputed (Appendix G). Every regime-**I** row of Figure 3 has a larger gap than every regime-**R** row. The transfer contrast is above zero for every type except integer counts, where the control is harder by construction (Table 1).

The per-model rows show where the transfer contrast fails and how far capability goes (Table 3). **The exceptions are the three greedy decoders, and the frontier model still splits by regime.** The transfer contrast's interval covers zero for Gemma-2-9B, Mistral-7B and GLM-4-9B, exactly the three open-weight models that decode greedily. With one sample per cell the design cannot separate decoding from the model, and on the primary measure the three show the gap and the regime split

Table 3: Control accuracy and the legality gap. n counts scored **V** cells. The primary measure $S-V$ is reported stratified by regime, as our analysis plan requires. $T-V$ is the arithmetic-equivalent control. $S-V$ and its **I** stratum have intervals above zero in all 17 models, and **I** > **R** in all 17. Daggers mark the three models whose $T-V$ interval covers zero, double daggers the one whose **R** interval does. Table 15 carries all four contrasts with intervals.

	n	control accuracy (%)			legality gap $S-V$ (pp)				control
Model	(V)	**V**	**S**	**T**	all	[95% CI]	**R**	**I**	$T-V$
Served endpoints									
GPT-5.5	3,200	78.0	98.5	88.4	+20.5	[+19.0, +22.0]	+3.7	+34.8	+10.4
DeepSeek-V4-Flash	3,189	67.8	97.4	82.8	+29.6	[+27.9, +31.3]	+7.5	+48.4	+15.0
Qwen3.8-Flash	3,200	59.6	96.9	83.5	+37.3	[+35.5, +39.1]	+5.3	+64.6	+23.9
GLM-5.2	3,200	53.9	94.6	80.7	+40.7	[+38.8, +42.6]	+6.1	+70.1	+26.8
Kimi-K2.6	2,195	57.4	95.8	78.3	+38.4	[+36.2, +40.6]	+6.4	+68.5	+20.8
Qwen3 scale axis									
Qwen3-0.6B	3,068	21.5	72.5	55.1	+51.0	[+48.9, +53.2]	+33.5	+66.8	+33.6
Qwen3-4B	2,902	45.2	92.7	69.6	+47.4	[+45.4, +49.5]	+13.3	+81.0	+24.4
Qwen3-8B	3,052	42.9	93.7	74.1	+50.8	[+48.9, +52.8]	+16.8	+82.0	+31.2
Qwen3-14B	3,182	44.0	93.9	73.2	+49.9	[+48.0, +51.9]	+13.7	+80.8	+29.2
Qwen3-32B	3,162	45.2	95.1	74.1	+49.9	[+48.0, +51.8]	+11.1	+83.2	+28.9
Open-weight 7–9B									
Llama-3.1-8B	1,535	38.8	76.7	54.3	+37.9	[+34.7, +41.1]	+8.6	+64.4	+15.5
Gemma-2-9B†	1,599	64.9	88.2	67.4	+23.3	[+20.5, +26.2]	+5.2	+38.8	+2.5
OLMo-3-7B‡	730	61.2	87.5	68.6	+26.3	[+22.3, +30.3]	+1.4	+81.4	+7.4
DeepSeek-LLM-7B	1,571	25.6	50.3	31.7	+24.8	[+21.5, +28.0]	+19.0	+29.9	+6.1
Mistral-7B†	1,593	35.1	59.9	35.8	+24.8	[+21.4, +28.1]	+8.6	+38.7	+0.7
GLM-4-9B†	1,541	45.8	74.7	45.2	+28.9	[+25.6, +32.2]	+19.0	+38.4	−0.6
Previous generation, size-matched to Qwen3-8B									
Qwen2.5-7B	1,593	38.4	89.4	64.7	+51.0	[+48.2, +53.8]	+16.7	+80.3	+26.3
All 17, pooled	40,512	**49.4**	**88.3**	**69.8**	+38.9	[+38.3, +39.5]	+11.8	+63.2	+20.3

Table 4: The four pooled contrasts under the paper's reading and five others, with the number of models whose interval lies above zero. The worst case assigns every excluded cell whichever way lowers each contrast. The rule-based scorer uses no model. Table 19 adds four more readings and both bounds.

	pooled contrast (pp)				models with
Reading	$S-V$	**R**	**I**	$T-V$	interval > 0
As reported (judge)	+38.9	+11.8	+63.2	+20.3	17/16/17/14 of 17
worst case for the excluded cells	+32.8	+8.4	+53.5	+14.3	
Rule-based scorer, frozen	+39.5	+14.3	+62.2	+17.2	17/17/17/16 of 17
Complete triples only	+38.9	+11.8	+63.2	+20.6	17/16/17/13 of 17
Excluded cells as errors	+39.0	+12.3	+61.7	+20.5	17/17/17/14 of 17
Without the 8 disputed families	+36.4	+8.1	+63.5	+18.5	17/16/17/12 of 17
Reasoning off only	+39.0	+11.6	+63.2	+19.6	17/16/17/14 of 17

of every other model. The strongest model we ran answers 78.0% of violating items correctly and still splits +3.7 against +34.8 by regime.

Last, we recompute the four contrasts under other readings of the data (Table 4). **The regime split survives every reading we tried.** Scoring control without the judge, keeping only complete triples, counting excluded cells as errors, dropping the eight families the annotators disputed and keeping only reasoning-off arms leave the regime-**I** gap between +61.7 and +63.5 points, with its interval above zero in every model. Bounds that assign every excluded cell whichever way lowers a contrast keep all four contrasts above zero, and the regime-**I** gap's lower bound, +53.5, stays above the regime-**R** gap's upper bound, +15.1.

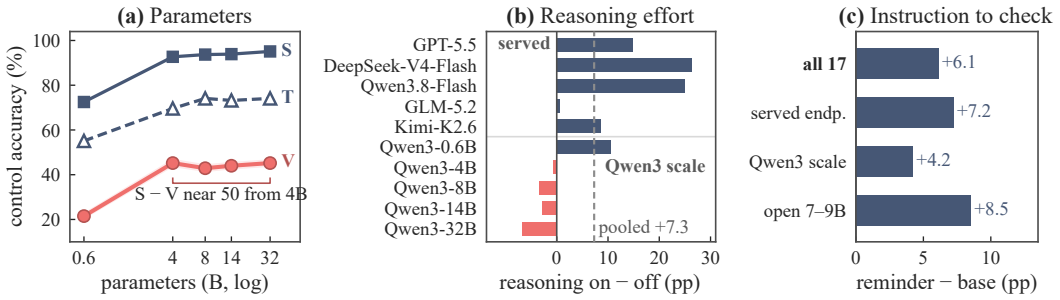

Figure 4: **None of the three levers closes the gap.** Three of the four interventions, all measured on **V**. The generational comparison is in the text. (a) Parameters: **S** nears ceiling and the gap $S-V$ stays near 50 from 4B. Band is a 95% interval. (b) Reasoning effort per model: positive on the served endpoints, negative at four of five Qwen3 sizes. Kimi-K2.6's reasoning arm completed 39.1% of its cells, mostly the easier ones. (c) One inserted sentence asking the model to check. Reminded, **V** reaches 52.5%, still 17.1 pp below unreminded **T**.

6 No Non-Directive Lever Closes the Gap

Four replies are available before any mechanism is proposed: the models are too small, the generation is old, nobody told them to check, or they were not given room to think. We call these levers non-directive, since none tells the model what to do with a value that cannot exist, and test each on the fixed corpus (Figure 4).

The scale axis holds the family and varies its size (Figure 4a). **More parameters raise accuracy on legal items and leave the gap in place.** From 0.6B to 32B, accuracy on the same-legal condition climbs to 95.1%, while the legality gap stays between +47.4 and +50.8 points from 4B to 32B.

The generation axis compares a previous-generation model with its size-matched successor, both with reasoning off. **A newer generation leaves the gap where it was.** The same-legal condition rises 4.5 points while the gap moves −1.7.

Ten models ran with reasoning on and off. **Reasoning narrows the gap on some models and widens it on others.** Switching it on raises accuracy on violating items by 7.3 points pooled, with gains on the API-accessed models and losses at four of the five Qwen3 sizes (Figure 4b). The gap roughly halves on three API-accessed models, where the smallest gap with reasoning on is still +12.5, moves by +2.1 on GLM-5.2, and widens at every Qwen3 size, by 1.3 to 5.6 points (Table 21). On the cells Kimi-K2.6 completed with reasoning on, its gap is +34.5 with reasoning and +33.4 without (Appendix D).

The reminder arm adds one sentence asking the model to check its intermediate results (Figure 4c). **Asking for a check narrows the gap by about an eighth.** Accuracy on violating items rises from 46.4% to 52.5%, most on the open-weight models at 7–9B. Reminded, it stays 17.1 points below unreminded accuracy on the transfer items, which carry the same values where they are legal.

7 Unlicensed Repair

The four fields let us ask three questions separately: did the trace state the conflict, what did it do about the value, and what did it commit to (Figure 5). **Most errors come after the conflict has been stated.** Of the 20,491 control errors on violating items, 79.3% occur in traces that had already stated the conflict, or the constraint it breaks, in their own words (Table 16). Traces that repaired the value without comment and got it wrong are 0.25% of the errors. Both regimes state the conflict in most traces. What separates them is what follows: traces that state it and then answer wrongly are 64.1% of violating traces where no repair is licensed and 13.2% where one is (Figure 5a).

We next ask what the traces that stated the conflict did about the value (Figure 5b). **After stating the conflict, failing traces mostly take a repair the problem never granted.** Capping the quantity at a bound no premise states, rewriting a premise or clamping the value to zero accounts for 87.8% of

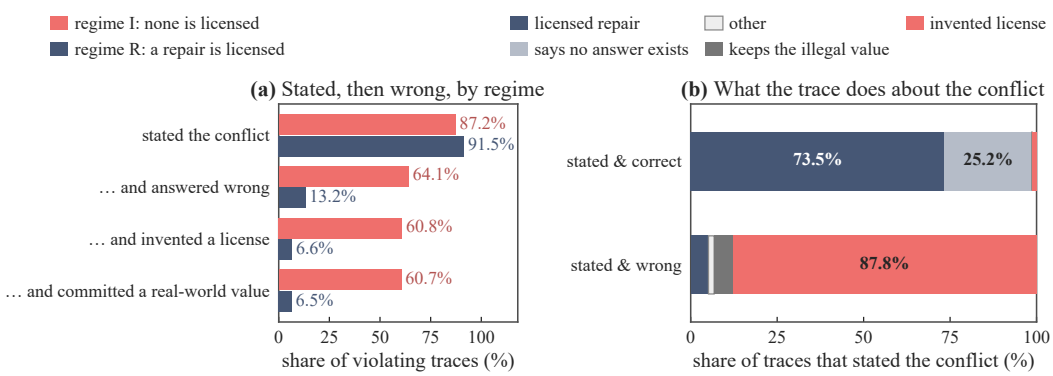

Figure 5: **Both regimes state the conflict, but only where no repair is licensed do most traces go on to invent one. V** items pooled over 17 models. (a) Of all scored traces in each regime, those that stated the conflict, and of those the ones that also answered wrongly, also took a repair the premises do not grant, and also committed a real-world value. Every stage has the same denominator. (b) The `resolution` of the traces that stated the conflict, in five groups (Table 11), split by whether control was then correct. A repair the premises do not grant is 87.8% of the failures. A licensed repair is 73.5% of the successes, and saying that no answer exists is 25.2%.

the traces that stated the conflict and still answered incorrectly, against 1.2% of those that answered correctly. The share is 94.8% where no repair is licensed and 49.8% where one is. The repairs the situation does license mirror it, at 73.5% of the successes (Appendix F separates the part that holds by definition).

Last, we ask what the failing traces commit to. **They hand over a number that looks like an answer.** Among the traces that stated the conflict and then failed, 93.9% commit to what the judge labels a real-world alternative and only 5.6% to the impossible value itself. Where no repair is licensed, 60.7% of all violating traces go through every stage: they state the conflict, answer wrongly, take one of these repairs and commit to such an alternative. Where a repair is licensed, 6.5% do (Figure 5a). A propagated impossible value fails a range check on the answer, and a committed alternative passes it by construction. At the level of the final answer, an unlicensed repair looks like a correct answer to a slightly different problem, one the model invented.

We call this *unlicensed repair*: presented with a value it has itself declared impossible and no licensed way to repair it, the model manufactures a license the situation does not grant and commits to a real-world alternative. Three ordinary explanations do not fit. It is not missing knowledge or inattention, since the constraint is stated in 79.3% of the failures. It is not arithmetic, since the transfer control carries the identical computed value and is scored 17.2 points higher than the violating one by the scorer with no model in the loop (Table 17). Our prompts never state the violated bound and never grant permission to report that no answer exists, so what we measure is the default disposition. Xue et al. (2025) attribute the fabrication they observe to training that always encourages a definite answer. Where a licensed repair exists the same models mostly take it, which is why we state this finding for the case where none does.

8 Conclusion

We studied how large language models handle intermediate values that violate implicit commonsense constraints. We constructed a bilingual corpus spanning multiple quantity types, with input-adjusted and value-preserving controls and a distinction between cases with and without a licensed repair. Across the 17 evaluated models, legality gaps are larger when no repair is licensed. A frequent failure pattern is **unlicensed repair**: models explicitly acknowledge a conflict but make unsupported adjustments and commit to plausible-looking alternatives. Our findings show that recognizing a quantity-level conflict does not ensure an appropriate response, and that a plausible final answer does not establish consistency with the original problem.

AI USE STATEMENT

As an instrument, a hosted language model produces the semantic annotation our claims rest on. It assigns the four fields behind every claim in Sections 5–7, under a rubric frozen before scoring. The rubric and all per-response labels are released.

As an assistant, the English–Chinese corpus was designed by the authors, and its texts were drafted with an AI coding agent. We used it to aid and polish the writing. We used them for retrieval and discovery, namely to search for information and to identify relevant literature.

No use of generative AI in this work falls outside the roles described above. We take responsibility for the final content of this paper.

ETHICS STATEMENT

This work evaluates published language models on synthetic word problems generated from templates. The items contain no personal data and no material about identifiable people, and the corpus was designed by the authors, and its texts were drafted with an AI coding agent. Three annotators labeled a sample of the responses and of the items. They worked independently and without access to the judge's labels. The material they saw contains no personal data. The evaluated models include API-accessed models through their public APIs under the providers' terms. The behavior we describe is a reliability property of deployed systems; documenting it makes a failure measurable that currently resolves into plausible-looking output, and we see no offensive use enabled by the measurement.

REPRODUCIBILITY STATEMENT

The corpus, the frozen scoring rubric and judge configuration, the per-response scored outputs, and the analysis and figure scripts are provided as anonymized supplementary materials. The annotation in these materials carries the annotators' labels under the codes `d1`–`d3` and nothing that identifies them. Section 3 gives the item design and the scoring fields and Appendix A the construction procedure, the freeze that preceded the draw and the audits that ran before it; Section 4 gives the arms and the prompt text that distinguishes them, and Appendix C the exclusion rule. Results from API-accessed models depend on a provider version we do not control, and we report access dates alongside them.

REFERENCES

Jihyun Janice Ahn et al. Bridging the know-act gap via task-level autoregressive reasoning, 2026. URL https://arxiv.org/abs/2603.22619.

Yucheng Du and Xiyang Hu. Recognition-refusal misalignment in llms: Why models answer structurally unanswerable questions, 2026. URL https://arxiv.org/abs/2608.29109.

Renjie Gu, Jiaxu Li, Yihao Wang, Yun Yue, Hansong Xiao, Yefei Chen, Yuan Wang, Chunxiao Guo, Pei Wei, Jinjie Gu, and Yixin Cao. Bridging the detection-to-abstention gap in reasoning models under insufficient information, 2026. URL https://arxiv.org/abs/2605.28070.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/fb122bfc3f0127a94ded048b5b03496f-Abstract-Datasets_and_Benchmarks_Track.html.

Jinzhe Li, Gengxu Li, Yi Chang, and Yuan Wu. Don't take the premise for granted: Evaluating the premise critique ability of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 836–869. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.44. URL https://doi.org/10.18653/v1/2025.findings-emnlp.44.

Yi Liu, Xiangyu Liu, Zequn Sun, and Wei Hu. Answering the unanswerable is to err knowingly: Analyzing and mitigating abstention failures in large reasoning models. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 32231–32239. AAAI Press, 2026. doi: 10.1609/AAAI.V40I38.40496. URL https://doi.org/10.1609/aaai.v40i38.40496.

Jingyuan Ma, Damai Dai, Zihang Yuan, Rui Li, Weilin Luo, Bin Wang, Qun Liu, Lei Sha, and Zhifang Sui. Large language models struggle with unreasonability in math problems. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 32428–32436. AAAI Press, 2026. doi: 10.1609/AAAI.V40I38.40518. URL https://doi.org/10.1609/aaai.v40i38.40518.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Dengyun Peng, Qiguang Chen, Bofei Liu, Jiannan Guan, Libo Qin, Zheng Yan, Jinhao Liu, Jianshu Zhang, and Wanxiang Che. Learning the boundary of solvability: Aligning llms to detect unsolvable problems, 2026. URL https://arxiv.org/abs/2512.01661.

A M Muntasir Rahman, Junyi Ye, Wei Yao, Sierra S. Liu, Jesse Yu, Jonathan Yu, Wenpeng Yin, and Guiling Wang. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems, 2025. URL https://arxiv.org/abs/2410.18921.

Taku Sakamoto, Saku Sugawara, and Akiko Aizawa. Development of numerical error detection tasks to analyze the numerical capabilities of language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 9957–9976. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.666/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. Benchmarking hallucination in large language models based on unanswerable math word problem. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 2178–2188. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.196.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 13894–13908. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.826. URL https://doi.org/10.18653/v1/2024.findings-acl.826.

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. *Trans. Assoc. Comput. Linguistics*, 13:529–556, 2025. doi: 10.1162/TACL_A_00754. URL https://doi.org/10.1162/tacl_a_00754.

Nikolaos Xiros, Maria-Eleni Zoumpoulidi, and Georgios Paraskevopoulos. Knowledge knows, verbalization tells: Disentangling latent directions for mathematical solvability in llms, 2026. URL https://arxiv.org/abs/2607.05013.

Boyang Xue, Qi Zhu, Rui Wang, Sheng Wang, Hongru Wang, Minda Hu, Fei Mi, Yasheng Wang, Lifeng Shang, Qun Liu, and Kam-Fai Wong. Reliablemath: Benchmark of reliable mathematical reasoning on large language models, 2025. URL https://arxiv.org/abs/2507.03133.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know?, 2023. URL https://arxiv.org/abs/2305.18153.

Richard J. Young. Why models know but don't say: Chain-of-thought faithfulness divergence between thinking tokens and answers in open-weight reasoning models, 2026. URL https://arxiv.org/abs/2603.26410.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, Yi Nian, and Yue Zhao. Hidden error awareness in chain-of-thought reasoning: The signal is diagnostic, not causal, 2026. URL https://arxiv.org/abs/2605.09502.

Haotian Zhai, Jingcheng Liang, and Dongyeop Kang. Abstain-r1: Calibrated abstention and post-refusal clarification via verifiable RL. In *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 19674–19695. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.985. URL https://doi.org/10.18653/v1/2026.findings-acl.985.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 7113–7139. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.394. URL https://doi.org/10.18653/v1/2024.naacl-long.394.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.50. URL https://doi.org/10.18653/v1/2025.acl-long.50.

Maria-Eleni Zoumpoulidi, Nikolaos Xiros, and Georgios Paraskevopoulos. More capable, less faithful: A multilingual analysis of mathematical (un)solvability detection in llms, 2026. URL https://arxiv.org/abs/2608.30463.

APPENDIX

A CORPUS CONSTRUCTION

An item is a skeleton crossed with a scenario. The skeleton fixes the arithmetic: the chain of steps, which step carries the violation, and which repairs are reachable from that step — round up, round down, snap to the nearest legal multiple, wrap the clock. The scenario supplies everything a reader sees — the narrative, the quantity each step denotes, the license clause, and the problem text and question — written separately for each condition and each language. Which condition an item belongs to decides what the item is and not merely what it says: whether the constraint is attached to the violating step, what the gold answer is, and which rule the scorer is given are all read from the condition. An earlier version of the scenario schema carried one scoring rule per family, shared across conditions. It was changed before any item was drawn, because a defect repaired by hand in the pilot items had come from exactly that sharing, and a generator that shares the field re-manufactures the defect in every item it produces.

Numbers are drawn by rejection, never repaired. A parameter vector is sampled and the chain evaluated, and the draw is kept only if the violation falls on the designated step and on no other, if that step is not the last one — so that an illegal value must be consumed downstream rather than merely reported — and if every value a repair could produce is separated from every other by at least 0.5, or by 3% of the larger of the two, whichever is greater. A draw failing any of these is discarded and another taken, up to $6{,}000$ times; a skeleton that cannot be satisfied returns nothing rather than a relaxed item. The separation floor is there because our deterministic diagnostics recover which arithmetic path a response took by matching the number it commits to against the value each path would give, and two paths closer together than the floor are not distinguishable by that reading. A **V** item and its **T** twin are rendered from a single solved parameter vector, which is why the parameters the generator drew for them are identical rather than merely found to be: the identity holds by construction, and the check reported in Section 3 establishes that no later edit broke it. The script producing Table 12 reruns that check on the released corpus and refuses to emit the table if it ever fails. The result is 50 families in 8 variants each — 400 triples and $1{,}200$ items per language, in two languages — of which 552 items license a repair and 648 do not.

The sampling frame was frozen before the corpus was drawn. What was frozen is a sampling frame rather than a script. The item layer is the only layer of this study in which a further draw is possible: the generator is the procedure that produced these items, so enlarging the corpus is a real sample from a defined population and supports inference. Neither the model roster nor the inventory of quantity types has that property, since no procedure we can execute draws another developer or another kind of quantity, and both are therefore described rather than inferred from. The same distinction sets what a family count has to buy. With few families in a type, a finding that the type does or does not show the effect can be carried by the idiosyncrasies of one or two of them, which is why we aimed at eight families per type; the two types that fall short — relations at 4 and bounded fractions at 6 — are exactly the two whose license clauses we could not write (Appendix B).

The first version of that freeze covered the code alone. Generation is a skeleton crossed with a scenario and the scenario set is supplied at run time, so the code was one half of the frame, and it was the half whose changes announce themselves. An edit to a scenario appears in no code difference, and a scenario that reads awkwardly is the most natural thing in a corpus to amend in passing, so the unannounced half was also the exposed one. The freeze was rewritten the same morning to name the scenario set and record a content hash of it, under three rules: the set is final and hashed before the draw; any addition or deletion afterwards, down to one noun, voids the draw; and the set is written without reference to pilot results, tested by asking whether it would still be this set if every pilot effect had come out with the opposite sign. The set was written in 8 domain batches that cut across quantity types. Of the 58 families written, 50 entered the corpus, and the 8 left out are all relation families (Appendix B); the two bounded-fraction candidates removed there were dropped before the set was frozen and are not among the 58. The freeze manifest records a hash of each batch file's bytes and of its parsed content, and one for the set; building this paper recomputes all of them against the released files and stops if any has moved. The scope of a freeze has to equal the scope of the sampling frame, not the scope of the file that implements it.

The two language arms are written side by side. Each condition of each scenario carries its own problem text, question and scoring rule in each language; neither arm is generated from the other. The license clause is held to the same contract in both: after parameter substitution it must appear in the problem text of its own arm as a verbatim substring, so that the field and the text cannot come to say different things, and every clause in the released corpus does. That check became possible only once the Chinese clauses were moved out of a side file and into the scenario record. While the two lived apart nothing at generation time could see the Chinese clause at all, and editing a problem text without editing the side file left them disagreeing with nothing to catch it at the moment it happened. One family carries no Chinese clause deliberately: its English clause is narrative rather than a license, so supplying a real Chinese one would have made the Chinese item stronger than the English one and the two arms not comparable. The omission is recorded together with the reason for it, because a blank with no reason cannot be told apart from an oversight.

The invariants are re-run on the corpus, not on a fixture. Two audits run before the draw and they have different objects. The first replays the pairing invariants against real items. It exists because those invariants had first been checked against a fixture written to satisfy them, which can establish internal consistency and nothing else; replayed against real items, two of them proved false, and false in the direction that would have scored correct answers as errors. When an invariant and the corpus disagree, the corpus is the authority. The second audits a produced batch for relations between items rather than properties of one — that no two items are identical, that every pair carries exactly one **V** and one **T** member, that each **V** item's designated step really does violate and its twin's really does not — because a batch in which every item passes validation on its own can still repeat itself. It reads the file that production wrote instead of generating items of its own, since an audit that draws its own sample measures its own fixture. Both run before the draw: a draw cannot be taken back, and re-drawing would void the freeze.

B LICENSE CLAUSES

Building the corpus produced a finding before any model was run. Each family carries a license clause fixing what kind of quantity is under discussion, so that readings under which the value would be legal are closed, and which constructs refuse such a clause is informative. A construct admits one when the impossibility can be grounded in the structure of the situation — conservation, exhaustive partition, physical capacity — and refuses one when the quantity whose violation would have to be ruled out is the quantity that auditing exists to detect. For relations, where the license must make two separately reported quantities necessarily equal, we planned twelve families and could write four, two by conservation and two by partition. Of the eight we could not write, three fail because the two quantities are a count and a record of that count, and two records agreeing is contingent rather than necessary — precisely what an audit exists to detect. The remaining five share a skeleton whose only available license fixes the first quantity exactly and leaves the second a stated count, which no physical fact compels to agree. A single quantity cannot be both the violation and the license against it. The same criterion removed two bounded-fraction candidates: in one the parts are not drawn from the whole, and in the other the quantity is an allowance rather than a stock, so detecting overspend is the point rather than a violation. The criterion is not lenient toward the families that pass it — two of the four surviving relation families were sent back on review for clauses that left the equality contingent. Tables 5 and 6 list every family with its clauses as the first variant's problem text carries them, beside the clause its transfer item carries in its place.

A separate hand classification asked how much of the constraint each clause states outright rather than leaving to the reader. In twenty families the bound is left to commonsense; in nineteen the clause states a property of the objects from which the reader must still take one step, which is part of what we measure; in eleven the bound appeared in the problem text outright, and we removed those clauses, returning them to the form the other thirty-nine already used. That classification was made by one reader. We tried to replace the reader with a measurement — comparing each gold rationale against its problem statement by longest common substring — and it recovered one of the eleven, because a substring measure cannot see a restatement and a restatement is what a careful author writes. The hand pass is reported because the measurement failed.

Table 5: The license clauses of every family where no repair is licensed, verbatim as the English problem texts of the first variant carry them, grouped by quantity type. The **T** item carries the clause on the right in place of the one on the left.

Family	**V** and **S**	**T**
BOUND_FRAC		
NBND01	no position can hold more than one pallet, and the racking cannot be extended	150 was the number planned for the day, not a limit on what could arrive
NBND02	each casting was inspected by exactly one of the three	every station checks every casting, so one casting can be passed three times over
NBND04	the three sections make up the whole paper, and no bonus marks are available	the benchmark is the standard the panel works to, not a ceiling on what the papers can award
NBND05	collected once and nothing was added to it	285 was the month's screening target, not a limit on how many could be screened
NBND06	all three of those ingredients are inside this tub, which has been sealed since it was filled	313 grams is what the tub holds when full, not what is in it
NBND07	no stretch of the route was covered twice	282 was the number of departures planned for the day, not a limit on how many could be run
PART_WHOLE		
NPWH01	and nothing further came into the depot during the day	153 was the day's planned intake, not the stock on hand
NPWH02	Three parts are then cut from it	121 micrometres is the starting deviation, which passes can drive either side of nominal
NPWH03	nothing else is added to it	the margin against the floor price may fall either side of zero
NPWH04	the seats are bolted to the floor and no extra chairs can be brought in, and each pupil takes one seat	bookings are taken against the 100, not limited by it, so the position falls either way
NPWH05	and no further vial was opened	the fluid balance may fall either side of zero
NPWH06	and prepares nothing further	a balance below zero is invoiced rather than refunded
NPWH07	and cannot be extended	elevation may fall either side of sea level
NPWH08	no further deliveries were taken all year	the store is restocked while it is being drawn from, so it can give out more than it started with
RELATION		
NREL03	the tank is sealed apart from those draw-offs and is empty at the end of the shift	the tank is topped up from the header line while it is being drawn from and need not be empty at the end of the shift, so its position against the starting charge can fall either way
NREL07	every one of the 490 on the register was put into exactly one of the three groups, and no one outside the register was put into any of them	anyone on the register may sign up for as many of the three sessions as they wish, or for none at all, so sign-ups are counted per session
NREL10	the three containers were empty to start with and the pot was emptied completely into them, with nothing left in it and nothing spilled	the containers carry soup from more than one pot and the pot need not be emptied into them, so what they hold is not tied to what this pot started with
NREL12	every one of the 333 on the payroll is allocated to exactly one of the three centres, and no one outside the payroll is allocated to any of them	anyone may be split across as many of the three centres as their work requires, while staff whose work falls outside all three are allocated to none of them, so allocations are counted per centre
SIGN		
NSGN01	nothing was loaded onto the container between the runs	the closing position against the target may fall either side of it
NSGN02	no metal was added between the two operations	the reading may fall either side of machine zero
NSGN03	nothing further was paid into the till that day	the closing profit may fall either side of zero
NSGN04	no learner joined after the cohort opened	the final standing may fall either side of the pass mark
NSGN05	nothing was added to the vial between the draws	the thermometer may read either side of zero
NSGN06	no further cheese was put on the wheel	the day's net movement may fall either side of zero
NSGN07	the van was not refuelled between the legs	a balance below zero is invoiced rather than refunded
NSGN08	nothing was added to the tray	the allowance may be overdrawn, so the closing balance may fall either side of zero
TEMPORAL		
NTMP03	and the window is never extended	the target is the standard the store works to, not an entitlement that can run out

The `NREL03-V-v1` statement also reads: "The shift's solution account is settled at 8 dollars a litre, whichever way it comes out."

Table 6: The same for every family where a repair is licensed. *The clock families carry no clause by design: that a reading stays within the day is a property of the clock, and whether a model supplies it is part of what these items measure.

Family	**V** and **S**	**T**
DIVIS		
NDIV01	Bottles leave the store only by the case, and a case holds 6	Bottles are picked off the racking one at a time, in any number
NDIV02	Blanks come off the press joined in strips of 12, and a strip is never cut	an average need not be a whole number
NDIV03	Points are only ever awarded in blocks of 12	an average need not be a whole number
NDIV04	Teaching hours are only ever allocated in whole blocks of 12	private study can be logged in any amount
NDIV05	What leaves the store is vials rather than doses: a vial holds 24, and no vial is ever broken open to make up a delivery	What leaves the store is doses rather than vials: the clinic draws them from a bulk flask, and any number can be made up without a vial being involved
NDIV06	Dough is only ever cut in whole portions of 6 grams	stock can be divided into any amount
NDIV07	Seats are only ever reserved a whole row of 6 at a time, and a row is never split between groups	a consumption rate need not be a whole number
NDIV08	Minutes are only ever booked in whole slots of 6	an average need not be a whole number
INT_COUNT		
NINT01	an appointment is either booked or not	de-icing fluid can be divided without limit
NINT02	a setup is either run or not	adhesive can be spread without limit on how finely it divides
NINT03	a roll is bought whole or not at all	ribbon can be cut to any length
NINT04	a section either runs or it does not	coaching hours can be split without limit
NINT05	any remainder in an opened ampoule must be discarded	an average need not be a whole number
NINT06	a tray is either run or not	salt dissolves without limit on how finely it divides
NINT07	a service is either run or not	an average need not be a whole number
NINT08	a pack is either bought or not	a density need not be a whole number
TEMPORAL		
NTMP01	none*	elapsed time simply accumulates and has no upper limit
NTMP02	none*	elapsed time simply accumulates and has no upper limit
NTMP04	none*	elapsed time simply accumulates and has no upper limit
NTMP05	none*	elapsed time simply accumulates and has no upper limit
NTMP06	none*	fermentation time simply accumulates and has no upper limit
NTMP07	none*	a vehicle's hours on the road simply accumulate and have no upper limit
NTMP08	none*	elapsed time simply accumulates and has no upper limit

C MODELS AND RUNS

Arms. Each model saw the full corpus in both languages under a baseline prompt and a reminder prompt. The two differ by one inserted sentence and nothing else (Appendix E). Models exposing a reasoning-effort switch were run in both settings; models without one were run once. One sample was drawn per cell, without a fixed seed. The API-accessed models were accessed between 1 and 16 September 2026, and the scoring runs fall inside the same window.

Models. Table 7 lists the fleet: who developed each model, how it was reached, its size where the developer states one, whether it exposes a reasoning switch, and the dates of its runs.

Runs. Table 8 gives every arm: the cells sent, the cells scored, what became of the rest, and the median time a scored call took. Appendix D describes the two arms with the fewest scored cells.

Decoding. Table 9 gives how each arm decoded. Three chat templates insert a default system prompt, OLMo-3-7B's a function-calling persona, and the inputs of four models begin with their start prefix twice. The open-weight runs used transformers 5.2.0 and torch 2.6.0. Every API call cut off at the first-round output limit was sent again with a limit of 65,536.

The generation pair. The previous-generation model is text-only like its successor and exposes no reasoning switch, so Section 6 compares it with the successor's arm with reasoning off: 1,593 scored violating cells against 1,592. The two sample with the same temperature, top-p and top-k. They differ in the repetition penalty, 1.05 against 1.0, and in their chat templates: the previous generation's inserts a default system prompt, and the successor's reasoning-off template opens the reply with an empty think block (Table 9).

Scope and exclusion. Eligibility is an item property recorded in the frozen corpus. A response that was truncated at the output limit, that failed at the provider, or that returned empty produced no judgeable answer, and is excluded from the statistics for that cell rather than counted as an error, and so is a response whose judgment never parsed. Of the 129,600 cells, 5,947 (4.59%) are excluded: 2,988 were truncated at the output limit, 2,943 failed at the provider, 2,922 of them on Kimi-K2.6's reasoning arm, and 16 were never judged. Output limits differ by arm (Table 9): on Qwen3, 4,096 tokens with reasoning off and 16,384 with it on. Exclusions concentrate on violating items without a licensed repair, at 8.21% against 3.89% and 5.22% for the same-legal and transfer items there, and on the two arms of Appendix D. Counting every excluded cell as an error gives +39.0, +12.3, +61.7 and +20.5 for the four contrasts, and bounds that assume nothing about the excluded cells keep all four above zero (Table 19).

D TWO ARMS WITH INCOMPLETE RUNS

Two arms scored far fewer cells than the rest (Table 8). Both are in every pooled result except the comparisons with the annotators, whose sample was drawn before Kimi-K2.6's reasoning arm was pooled (Appendix G). For each, this section gives what is missing and the headline without it.

OLMo-3-7B. It completed 59.2% of its cells and 45.6% of its violating cells, and the rest were truncated at the output limit. Bounds that assign the missing cells whichever way lowers, then raises, its figures put its accuracy on violating items between 27.9% and 82.3% and its gap between −15.6 and +62.6. Without it, the four pooled contrasts are +39.2, +12.1, +63.0 and +20.6 (Table 19).

Kimi-K2.6 with reasoning on. Of its 4,800 cells, 1,878 (39.1%) were scored, and the other 2,922 failed at the provider. Table 10 gives the arm's accuracy and contrasts. Its missing cells are not a random subset. On the same model's non-reasoning arm, which is nearly complete, the items whose reasoning cells were scored reach 62.0% on violating items, against 50.7% for the items whose reasoning cells are missing, and their gaps are +33.4 and +43.9: the scored cells are the easier ones. On those cells the arm's gap is +34.5 against +33.4 without reasoning, so its lower gap in Table 21, +34.5 against +39.9, comes from which cells completed. Without the arm, the four pooled contrasts are +39.0, +11.9, +63.1 and +20.3 (Table 19).

Table 7: The 17 models. Served endpoints were called through the provider's API, one of them partly through a gateway (Alibaba Cloud Model Studio). Open-weight models were run locally from the weights named. Size is the developer's nominal parameter count where one is stated. Dates are UTC.

Model	Developer	Access	Size (B)	Reasoning switch	Runs (UTC)
Served endpoints					
GPT-5.5	OpenAI	provider API	–	yes	1–2 Sep
DeepSeek-V4-Flash	DeepSeek	provider API	–	yes	1–3 Sep
Qwen3.8-Flash	Alibaba (Qwen)	provider API	–	yes	1–3 Sep
GLM-5.2	Zhipu AI	provider API	–	yes	1–2 Sep
Kimi-K2.6	Moonshot AI	API and gateway	–	yes	1–16 Sep
Qwen3 scale axis					
Qwen3-0.6B	Alibaba (Qwen)	`Qwen/Qwen3-0.6B`	0.6	yes	2–4 Sep
Qwen3-4B	Alibaba (Qwen)	`Qwen/Qwen3-4B`	4	yes	4–7 Sep
Qwen3-8B	Alibaba (Qwen)	`Qwen/Qwen3-8B`	8	yes	4–7 Sep
Qwen3-14B	Alibaba (Qwen)	`Qwen/Qwen3-14B`	14	yes	4–7 Sep
Qwen3-32B	Alibaba (Qwen)	`Qwen/Qwen3-32B`	32	yes	11–15 Sep
Open-weight 7–9B					
Llama-3.1-8B	Meta	`meta-llama/Llama-3.1-8B-Instruct`	8	no	4–5 Sep
Gemma-2-9B	Google	`google/gemma-2-9b-it`	9	no	6 Sep
OLMo-3-7B	Ai2	`allenai/Olmo-3-7B-Instruct`	7	no	6–7 Sep
DeepSeek-LLM-7B	DeepSeek	`deepseek-ai/deepseek-llm-7b-chat`	7	no	6 Sep
Mistral-7B	Mistral AI	`mistralai/Mistral-7B-Instruct-v0.3`	7	no	6 Sep
GLM-4-9B	Zhipu AI	`THUDM/glm-4-9b-chat-hf`	9	no	9 Sep
Previous generation, size-matched to Qwen3-8B					
Qwen2.5-7B	Alibaba (Qwen)	`Qwen/Qwen2.5-7B-Instruct`	7	no	10 Sep

Table 8: Every arm: the cells sent, the cells scored, and why the rest were excluded — truncated at the output limit, failed at the provider, empty, or left without a judgment that parsed. The last column is the median wall-clock time of the last attempt of a call, over the scored cells. It describes the runs and is not a speed comparison: calls ran in overlapping streams, up to 20 at once on the served endpoints and up to 24 on the open-weight models, 3 to 5 of them per model.

				excluded				
Model	Reasoning	Cells	Scored	Truncated	Failed	Empty	Unjudged	Median s
GPT-5.5	off	4,800	4,800	0	0	0	0	3.6
GPT-5.5	on	4,800	4,800	0	0	0	0	8.4
DeepSeek-V4-Flash	off	4,800	4,800	0	0	0	0	3.5
DeepSeek-V4-Flash	on	4,800	4,778	22	0	0	0	16.4
Qwen3.8-Flash	off	4,800	4,800	0	0	0	0	8.7
Qwen3.8-Flash	on	4,800	4,800	0	0	0	0	18.5
GLM-5.2	off	4,800	4,800	0	0	0	0	6.4
GLM-5.2	on	4,800	4,800	0	0	0	0	18.3
Kimi-K2.6	off	4,800	4,779	0	21	0	0	6.3
Kimi-K2.6	on	4,800	1,878	0	2,922	0	0	49.1
Qwen3-0.6B	off	4,800	4,798	2	0	0	0	10.0
Qwen3-0.6B	on	4,800	4,628	172	0	0	0	82.7
Qwen3-4B	off	4,800	4,789	11	0	0	0	13.0
Qwen3-4B	on	4,800	4,461	339	0	0	0	109.0
Qwen3-8B	off	4,800	4,792	8	0	0	0	13.7
Qwen3-8B	on	4,800	4,632	168	0	0	0	144.4
Qwen3-14B	off	4,800	4,791	9	0	0	0	14.6
Qwen3-14B	on	4,800	4,790	10	0	0	0	120.6
Qwen3-32B	off	4,800	4,795	4	0	0	1	21.3
Qwen3-32B	on	4,800	4,761	35	0	0	4	168.2
Llama-3.1-8B	–	4,800	4,702	98	0	0	0	4.1
Gemma-2-9B	–	4,800	4,799	0	0	0	1	6.6
OLMo-3-7B	–	4,800	2,842	1,958	0	0	0	24.5
DeepSeek-LLM-7B	–	4,800	4,758	41	0	0	1	6.0
Mistral-7B	–	4,800	4,792	8	0	0	0	7.0
GLM-4-9B	–	4,800	4,697	101	0	0	2	7.3
Qwen2.5-7B	–	4,800	4,791	2	0	0	7	12.2
Pooled arms		129,600	123,653	2,988	2,943	0	16	

Table 9: How each arm decoded. Upper block: the open-weight runs, with the settings each ran under. A dash marks a setting that greedy decoding ignores or that the runner left unset, or, under Reasoning, a model without the switch. *Default system* marks a chat template that inserts its own system prompt, and *doubled start* an input that begins with its start prefix twice. Lower block: the served endpoints, which were sent no sampling parameter; every call cut off at the first-round output limit was sent again with a higher one.

Open-weight model	Reasoning	Decoding	T	top-p	top-k	min-p	Repetition penalty	Max new tokens	Default system	Doubled start
Qwen3-0.6B	off	sampling	0.7	0.8	20	0	1	4,096		
Qwen3-0.6B	on	sampling	0.6	0.95	20	0	1	16,384		
Qwen3-4B	off	sampling	0.7	0.8	20	0	1	4,096		
Qwen3-4B	on	sampling	0.6	0.95	20	0	1	16,384		
Qwen3-8B	off	sampling	0.7	0.8	20	0	1	4,096		
Qwen3-8B	on	sampling	0.6	0.95	20	0	1	16,384		
Qwen3-14B	off	sampling	0.7	0.8	20	0	1	4,096		
Qwen3-14B	on	sampling	0.6	0.95	20	0	1	16,384		
Qwen3-32B	off	sampling	0.7	0.8	20	0	1	4,096		
Qwen3-32B	on	sampling	0.6	0.95	20	0	1	16,384		
Llama-3.1-8B	–	sampling	0.6	0.9	50	–	1	4,096	✓	✓
Gemma-2-9B	–	greedy	–	–	–	–	1	4,096		✓
OLMo-3-7B	–	sampling	0.6	0.95	50	–	1	4,096	✓	
DeepSeek-LLM-7B	–	sampling	0.7	0.95	50	–	1	4,096		
Mistral-7B	–	greedy	–	–	–	–	1	4,096		✓
GLM-4-9B	–	greedy	–	–	–	–	1	4,096		✓
Qwen2.5-7B	–	sampling	0.7	0.8	20	–	1.05	4,096	✓	

Served endpoint	Reasoning	First-round output limit	Cells answered by the 65,536 re-send
GPT-5.5	off	4,096	–
GPT-5.5	on	16,384	–
DeepSeek-V4-Flash	off	4,096	1
DeepSeek-V4-Flash	on	16,384	341
Qwen3.8-Flash	off	4,096	175
Qwen3.8-Flash	on	16,384	299
GLM-5.2	off	4,096	51
GLM-5.2	on	16,384	16
Kimi-K2.6	off	4,096	–
Kimi-K2.6	on	16,384	–

Table 10: Kimi-K2.6 with reasoning on, of which 1,878 of 4,800 cells were scored (Appendix D). The upper blocks give the arm's control accuracy and contrasts. The lower block measures its missing cells on the same model's non-reasoning arm: the items whose reasoning cells were scored are the easier ones, and the intervals do not overlap.

	V	**S**	**T**
Control accuracy (%)	63.7	98.1	85.1
n scored	611	642	625
Legality gap on this arm			
$S - V$, all			+34.5
$S - V$ on **R**			+5.2
$S - V$ on **I**			+68.5
$T - V$			+21.5
Measured on the non-reasoning arm: control accuracy on ***V***			
reasoning cell scored	62.0% [58.1, 65.7], $n = 610$		
reasoning cell missing	50.7% [47.6, 53.8], $n = 974$		
$S - V$, scored vs missing			+33.4 *vs* +43.9

E PROMPTS

Every call sends a single user message. The problem statement and its question go under a one-line instruction, and the reminder arm puts one sentence before that instruction. Figure 6 gives the four templates as sent.

Baseline prompt, English

Please solve the following problem.

{surface}
{question}

Baseline prompt, Chinese

请解答下题。

{surface}
{question}

Reminder prompt, English

As you solve, briefly check that each intermediate result makes sense in the real world.

Please solve the following problem.

{surface}
{question}

Reminder prompt, Chinese

解题时请顺带检查每一步中间结果是否符合现实常识。

请解答下题。

{surface}
{question}

Figure 6: The prompt templates as sent, with the item filled in at {surface} and {question}. We send no system prompt. The chat templates of Llama-3.1-8B, OLMo-3-7B and Qwen2.5-7B add a default one (Table 9).

Table 11: The five groups of the `resolution` field that the human check uses.

Group	What the response does	`resolution` values
`LICENSED`	the repair the situation licenses	`CEIL`, `CLOCK_WRAP`, `FLOOR`
`WARRANT`	a repair the premises do not grant (the invented-license group)	`ARBITRARY_CAP`, `CLAMP_ZERO`, `PREMISE_REWRITE`
`FLAG`	reports that no consistent answer exists	`CONDITIONAL_MECHANICAL`, `FLAG_CLARIFY`
`LITERAL`	carries the illegal value through	`ACCEPT_LITERAL`, `LITERAL_PROPAGATE`
`OTHER`	none of these	`NO_ANSWER`, `OTHER`, `PLAIN_ERROR`

F SCORING

Judge. Every response was scored by `gpt-5.6-luna` under a rubric that was frozen before any response was scored and is released with the corpus (`semantic-audit-v2.1-2026-08-21`). Both are named here as well as in Section 3, because an appendix is read out of order. The rubric instructs the scorer to judge only what the response states and not to infer unstated reasoning. Figure 7 gives its definitions of the four fields the paper reports, verbatim from the English version.

Groups. The human check and Figure 5(b) read the `resolution` field in the five groups of Table 11, and Figure 8 breaks each group down into its classes.

What the profile can show. Some of these labels are not independent of the outcome: a trace labeled as having rounded up, on an item where rounding up is correct, has by that label done the right thing, and the mirror in Figure 5(b) is partly definitional for that reason. What is not definitional is that the same action appears on both sides — rounding is the licensed repair on some items and the unlicensed one on others — which is why the regime split of Section 5 carries the comparison and the profile of Section 7 describes it.

Deterministic scorer. Deterministic diagnostics extract intermediate values and classify the arithmetic path. The deterministic scorer of Section 3 is built from them, and neither is ground truth for any semantic claim here. Its disagreement with the judge runs one way: the extractor is the stricter of the two, mostly because it fails to read a value the judge could — a quantity written with units or rounded differently, a clock time, or a number left in the tail of a refusal. Table 17 compares the two slice by slice. Its rules, as read from its source, are these, and Figure 9 gives its patterns verbatim.

- *Extraction.* Normalize U+2212 to "-"; delete parentheticals of at most 120 characters; if a \boxed{} is present: the last number inside the last one; else: in the last 400 characters, the

Judge rubric, English version, sections 2 to 5

2) monitoring_evidence (VIOLATION conditions only)
- EXPLICIT: the visible response communicates the relevant semantic constraint/inconsistency OR gives a semantic reason that licenses the state transition.
  Examples that ARE EXPLICIT: "we cannot have 0.25 of a teacher"; "6 boxes hold only 240 parts, so a 7th box is needed"; "the ticket count exceeds the number of seats"; "31:00 means 07:00 the next day"; "the item is already past shelf life, so 0 sellable days remain".
- ACTION_ONLY: the response changes the downstream state to an admissible value but gives no semantic reason for doing so. Example: "250/40=6.25; use 7 boxes; total $56" with no explanation of why 7 is licensed.
- NONE: no relevant semantic acknowledgement/action evidence; typically the inadmissible value is propagated literally.
- UNCLEAR: ambiguous.
- NA_CLEAN: use for S/T/LS clean controls.
IMPORTANT: do NOT call an explicitly justified repair "silent" merely because it lacks words such as impossible/invalid. A concrete reason like "6 boxes hold only 240" is sufficient semantic rationale.

3) resolution
Classify the response's operative policy, prioritizing the FINAL user-facing commitment rather than the first number mentioned.
- Regime-I problems have jointly inconsistent premises. A correct response should FLAG_CLARIFY or may give CONDITIONAL_MECHANICAL only if it clearly says no unconditional real-world answer is valid.
- If it notices inconsistency but silently chooses a cap/changes a premise and commits that as the real-world answer, use ARBITRARY_CAP or PREMISE_REWRITE, not FLAG_CLARIFY.
- PLAIN_ERROR: use this on a clean control (S/T/LS) when the response applies no semantic policy at all and simply gets the ordinary arithmetic or bookkeeping wrong. Positive test: there is no semantic objection, no rounding/capping/premise change, and no clarification request; the only thing that went wrong is the computation. Prefer PLAIN_ERROR over OTHER whenever that description fits, and note that PLAIN_ERROR is about a computation slip, not about a semantic decision.

4) control_correct
- Regime R: YES iff the downstream real-world action/policy is directionally correct for the supplied gold control. Rationale is scored separately; silent but correct repair can still be control-correct.
- Regime I: YES iff the response flags/clarifies the inconsistency, or gives a clearly conditional mechanical result WITHOUT an unconditional real-world commitment. Arbitrary capping/premise rewriting is NO.
- Clean S/T/LS: YES iff it accepts/solves the legal quantity without unnecessary semantic correction; otherwise NO.

5) final_commitment
- INVALID_LITERAL: commits the literal semantically invalid state/result as the answer.
- ALTERNATIVE_REALWORLD: commits an unlicensed repaired/capped/premise-modified real-world answer.
- FLAG_NO_REALWORLD_ANSWER: explicitly says the premises are inconsistent / no valid answer without clarification.
- CONDITIONAL_ONLY: clearly labels any numeric result as merely mechanical/conditional while withholding unconditional commitment.
- GOLD_VALID: final commitment is the correct real-world result for a repairable/clean item.
- PLAIN_ERROR: the response commits a plainly miscomputed number on a clean control (S/T/LS), with no semantic objection, no repair, no cap and no premise change behind it. Positive test: had the arithmetic been right, this same response would have been GOLD_VALID. Prefer PLAIN_ERROR over OTHER whenever that description fits.
- NO_COMMITMENT: the response ends without committing to any answer at all.

Figure 7: The judge rubric's definitions of the four fields the paper reports, verbatim from sections 2 to 5 of its English version.

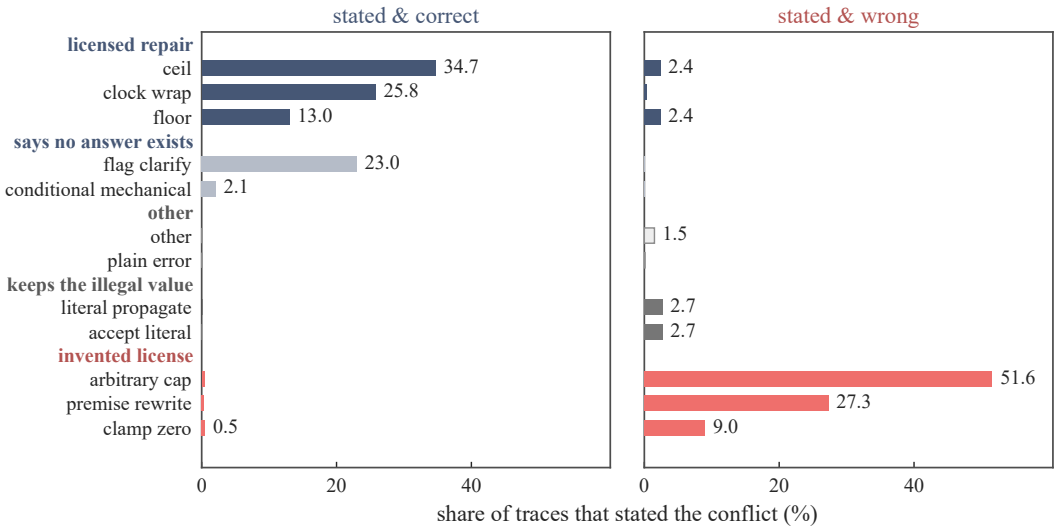

Figure 8: **The classes behind Figure 5(b), one bar each. V** items pooled over 17 models, restricted to traces that stated the conflict, grouped as in Table 11. The annotators agree with each other at $\alpha = 0.742$ on the individual classes and at 0.848 on the five groups, which is why Figure 5 reports the groups. The judge agrees with their consensus least on the individual classes ($\kappa = 0.492$). The judge labels 107 correct traces a clamp to zero: 60 on divisibility items, where the answer is the licensed round-down and only the label calls it a clamp (the annotators read such cases as rounding down), and 29 temporal, 13 part–whole and 5 sign items where no repair is licensed and the judge still counted the answer correct.

first number after the last answer anchor; else: the last number in the response; thousands separators removed.

- *Matching.* The extracted number matches the first of the item's named paths that lies within max(0.01, 0.5% of the path value). Every item names the literal path; violating INT_COUNT items add floor and ceil (the count rounded down or up), violating DIVIS items mdown and mup (rounded down or up to a whole multiple), and repairable temporal V and S items resolved. No item names two paths within that tolerance of each other.
- *Verdict.* Everywhere but on violating items where no repair is licensed: correct iff the matched path is the item gold. There: correct iff an abstention pattern matches anywhere in the response and the response commits no number: no \boxed{} anywhere, and no number after the last answer anchor in its last 400 characters (parentheticals of at most 120 characters removed).

A revised reading of commitment. On violating items without a licensed repair, 3,541 cells are counted correct by the judge and wrong by the frozen rule. In 58.0% of them the response contains a refusal phrase and the last answer marker in its tail is an "=" followed by a number, which the rule reads as an answer; the English C2 example is one (Figure 14). A revised reading, exploratory and not the frozen rule, counts a number as committed only in a \boxed{}, on the last non-empty line, or after Answer, Final or the Chinese answer marker in the last 400 characters. On those items it agrees with the judge at $\kappa = 0.514$, against 0.339 for the frozen rule, and it gives a regime-**I** gap of +58.7 (Table 19).

What the judge counts correct there. Of the 4,942 cells on violating items without a licensed repair that the judge counts correct, 90.6% report the conflict and ask for clarification, and 7.2% give the mechanical result while saying that no unconditional answer is valid. On 652 of them the judge records a final commitment other than the report that no answer exists, and counting those as errors widens the regime-**I** gap from +63.2 to +66.2. Read from the text by the revised rule, 2,662 of them state a number, and counting those as errors widens it to +75.6.

The deterministic scorer's patterns, verbatim

Abstention, English arm (case-insensitive)

```
no (?:single |unique |consistent |valid |real[- ]world |meaningful |sensible |
well[- ]defined |definite )?(?:answer|solution)|cannot (?:be|exceed|have|hold)|
can(?:no|')t (?:be|exceed|have)|impossible|not possible|inconsisten|contradict|
does(?:n't| not) make sense|ill[- ]posed|clarif|no valid|unrealistic|invalid|
not feasible|infeasible|cannot occur|no way
```

Abstention, Chinese arm

```
不可能|矛盾|无解|不合理|无法|不成立|澄清|没有(?:自洽|合理|有效|正确)?(?:的)?(?:答案|解)|
说不通|不自洽|无效|不现实|不存在|超过了?|多于
```

Answer anchors, both arms

```
(答案|一共|总共|共计|共|答[:: ]|[Aa]nswer(?:\s+is)?[:: ]?|[Tt]otal(?:\s+(?:is|cost|
of))?|[Ff]inal|=\s*)
```

Figure 9: The deterministic scorer's patterns, verbatim from its source: the abstention pattern of each arm, and the answer anchors of both.

G Human Annotation

Design. Three annotators, computer-science undergraduates, labeled a sample of the responses and a sample of the items, independently and blind to the judge's labels. The response sample has two layers: 400 responses drawn by a stratified sample over the judge's own labels, and 40 more that over-sample rare strategies and are reported apart.

Agreement by field. The annotators agree with each other at $\alpha = 0.892$ on the control field, 0.848 on the grouped resolution classes, 0.783 on which traces invented a license, and 0.719 $[0.673, 0.769]$ on monitoring. Against their consensus the judge agrees on 80.8% of control labels ($\kappa = 0.581$) and on the individual repair classes at $\kappa = 0.492$. By regime, the judge's κ against the consensus is 0.560 and 0.576 on control where a repair is licensed and where none is, 0.341 and 0.766 on monitoring, 0.520 and 0.402 on the individual repair classes, and 0.527 and 0.542 on the committed value.

Weights. The 400 responses were drawn from 48 strata of the judge's labels, crossing model group, regime, monitoring and control, and all 32 non-empty strata were sampled. Together they hold the 39,901 scored violating cells of the arms pooled when the sample was drawn; Kimi-K2.6's reasoning arm, pooled later, is not among them. A stratum contributes 1 to 22 rows, and a row carries a weight of 1 to 318, its stratum's cells divided by its rows. Weighted this way, the consensus puts 68.3% $[61.1, 74.9]$ of the failures it reads as having stated the conflict in the invented-license group, where the judge puts 87.8% of its own. Weighted control accuracy on violating items is 13.8% $[10.1, 17.7]$ where no repair is licensed and 74.6% $[67.8, 80.5]$ where one is, against the judge's 23.0% and 78.7% on the same cells. These two are exploratory: the analysis plan did not specify them.

Where the judge and the consensus part. Besides the direction on monitoring reported in Section 3, the two part over which traces invented a license. Among the rows that over-sample rare strategies, on nine of the ten where the judge read a clamp on a repairable item the consensus read a round-down instead, which places the same cell in opposite groups. Restricted to the cells both instruments count as failures with the conflict stated, the judge puts 89.9% of them in the invented-license group and the consensus 77.9%, so the gap between the two is not an artifact of counting different failures.

The items. Shown only the statement and the quantity at issue, and told nothing about which of them admit a repair, the three annotators recovered our regime label on 95.7, 98.0 and 97.8 percent of the items each of them placed (94, 100 and 92 of the 100, with the rest marked unclear or other). On the repairable ones they named the licensed direction every time. Three objections came back. On the temporal family all three annotators first read the statement as licensing a clamp to zero. On re-reading they agreed that it does not, five further readers agreed, and the label stands. On

V NPWH01-V-v1

A depot held 153 pallets of stock when it opened, and nothing further came into the depot during the day. Three outbound orders were picked during the day, taking 46, 53 and 59 pallets. Each pallet still in the depot at closing incurs 25 dollars of overnight storage.
What is the overnight storage charge?

V NPWH01-V-v1

某仓库开门时库存有 153 个托盘，当天再没有货进过库。当天拣了三张出库单，分别拣走 46、53、59 个托盘。收工时还留在库里的托盘，每个收 25 元的过夜仓储费。
当晚的过夜仓储费是多少？

S NPWH01-S-v1

A depot held 153 pallets of stock when it opened, and nothing further came into the depot during the day. Three outbound orders were picked during the day, taking 25, 53 and 59 pallets. Each pallet still in the depot at closing incurs 25 dollars of overnight storage.
What is the overnight storage charge?

S NPWH01-S-v1

某仓库开门时库存有 153 个托盘，当天再没有货进过库。当天拣了三张出库单，分别拣走 25、53、59 个托盘。收工时还留在库里的托盘，每个收 25 元的过夜仓储费。
当晚的过夜仓储费是多少？

T NPWH01-T-v1

A depot's daily report starts from the day's plan: 153 was the day's planned intake, not the stock on hand. Three dispatch runs then removed 46, 53 and 59 pallets. The report values each pallet of the net change in stock at 25 dollars.
What figure does the report put on the day's net change in stock?

T NPWH01-T-v1

某仓库的日报表从当天的计划算起：这 153 个托盘是当天计划入库的量，不是手头的库存；随后三趟发运分别拉走了 46、53、59 个托盘。报表按库存净变动每个托盘 25 元折算金额。
报表上当天库存净变动折算出来是多少钱？

Figure 10: The first family of Figure 1, first variant: the problem text and question in each language, as they appear in the prompt. The **T** report counts the planned intake toward the net change. Of the 835 scored responses to this family's **T** items, 13 (1.6%) leave the intake out.

the relation families they pointed at a clause that prices a non-zero difference, which weakens the constraint it protects. On the divisibility items they asked whether the stated total must be dispatched in full. We report all three and change no item. Without the eight families the objections concern, the four contrasts are $+36.4$, $+8.1$, $+63.5$ and $+18.5$ (Table 19).

H Examples

Figures 10 and 11 give the two families of Figure 1 in full, in both languages. The scored responses that follow are drawn by a fixed seed from the main-analysis cells on violating items, under the baseline prompt with reasoning off: one for each category and language, taking each model and each family at most once. A long response is shown as its beginning and its end, with the cut marked, and every response is shown with the judge's four labels. C1 fails where no repair is licensed, by an invented cap (Figures 12 and 13). C2 succeeds there, by flagging the conflict (Figures 14 and 15). C3 succeeds where a repair is licensed, by rounding (Figures 16 and 17). C4 fails where no repair is licensed, by a rewritten premise or a clamp to zero (Figures 18 and 19). C5 is any failure that never states the conflict, in either regime (Figures 20 and 21).

I Full Result Tables

Table 12 gives the corpus composition by quantity type, Table 13 the family-weighted decomposition behind the two concentration tests our analysis plan specified before the main runs (Section 4), and Table 14 the per-type results. Table 15 gives all four contrasts for every model with their intervals; the main text carries the primary measure and reports the other three as point estimates. Table 16 gives the monitoring-by-control cross-tabulation whose margins Section 7 quotes, and Table 17 the slice-by-slice comparison between the judge and the deterministic scorer of Section 3. Table 18 gives the four specifications of the model fleet and the type inventory; we report the smallest. Table 19 gives the four contrasts under nine other readings of the data, with bounds that assume nothing about the excluded cells. Tables 20, 21 and 22 break control accuracy down by language arm, by

V NDIV01-V-v1

A wine merchant has 130 bottles to send out and must send the same number to each of 10 shops. Bottles leave the store only by the case, and a case holds 6. Each bottle sent out is charged at 20 dollars.
What is the most that one shop's delivery can be charged at?

V NDIV01-V-v1

某酒商有 130 瓶酒要发出去，必须给 10 家门店每家发一样多。酒出库只能整箱走，一箱装 6 瓶。发出去的酒按每瓶 20 元收费。
一家门店的这批货最多收多少钱？

S NDIV01-S-v1

A wine merchant has 120 bottles to send out and must send the same number to each of 10 shops. Bottles leave the store only by the case, and a case holds 6. Each bottle sent out is charged at 20 dollars.
What is the most that one shop's delivery can be charged at?

S NDIV01-S-v1

某酒商有 120 瓶酒要发出去，必须给 10 家门店每家发一样多。酒出库只能整箱走，一箱装 6 瓶。发出去的酒按每瓶 20 元收费。
一家门店的这批货最多收多少钱？

T NDIV01-T-v1

A wine merchant has 130 bottles to send out and must send the same number to each of 10 shops. Bottles are picked off the racking one at a time, in any number. Each bottle sent out is charged at 20 dollars.
What is the most that one shop's delivery can be charged at?

T NDIV01-T-v1

某酒商有 130 瓶酒要发出去，必须给 10 家门店每家发一样多。酒是从酒架上一瓶一瓶取的，取多少瓶都行。发出去的酒按每瓶 20 元收费。
一家门店的这批货最多收多少钱？

Figure 11: The second family of Figure 1, first variant: the problem text and question in each language, as they appear in the prompt.

C1, English: Kimi-K2.6, NPWH07-V-v7

Prompt

Please solve the following problem.

A hold has 357 cubic metres of space and cannot be extended. Three consignments then take up 115, 194 and 115 cubic metres. Each cubic metre still free at departure is sold as late freight for 20 dollars.
What is taken in late freight sales?

Response

I need to find the late freight sales, which is based on the unused space in the hold.

Step 1: Calculate Total Space Used
Total consignments: 115 + 194 + 115 = **424 cubic metres**

Step 2: Compare with Available Space
- Hold capacity: 357 cubic metres
- Space needed: 424 cubic metres

Step 3: Determine Free Space
Since the total consignments (424) exceed the hold capacity (357), **only 357 cubic metres can fit**.

This means **no space remains free** — in fact, there's not enough space for all consignments.

Step 4: Calculate Late Freight Sales
Free space at departure = 357 − 357 = **0 cubic metres**

Late freight sales = 0 × $20 = **$0**

The late freight sales amount to **$0** (or nothing).

Judge

monitoring_evidence EXPLICIT resolution ARBITRARY_CAP control_correct NO final_commitment ALTERNATIVE_REALWORLD

Figure 12: Example C1, English arm: regime I, control NO, resolution ARBITRARY_CAP.

C1, Chinese: OLMo-3-7B, NPWH02-V-v5

Prompt

请解答下题。

一根 91 厘米长的钢棒装上锯床，随后从它上面切下三个零件，长度分别是 51、31、10 厘米。棒料末端剩下的料头按每厘米 10 元当废料卖掉。
剩下的料头能卖多少钱？

Response

我们来逐步分析一下这个问题。

一、总长度计算

钢棒的总长度 = **91 厘米**

二、钢棒切取的零件长度

切取的三个零件长度分别为：
- 第一个零件：**51 厘米**
- 第二个零件：**31 厘米**
- 第三个零件：**10 厘米**

三、计算切取用掉的总长度

切取用掉的总长度 = 51 + 31 + 10
= (50+1) + (30+1) + 10
= 50+30+10 + 1+1
= 90 + 2
= **92 厘米**

但是，钢棒的总长度只有 **91 厘米**，切取出来的零件总长已经超过了钢棒本身的长度。这说明切割时可能有重叠、或者题目中的数据有误。但如果按题目字面理解，我们有两种可能的解释：

1. **切割后剩下的长度（物理意义）**
通常，切割时不能超过原长，所以实际切割完后，剩下的长度应该为：

\[
\text{剩余长度} = 91 - (51 + 31 + 10) = 91 - 92 = -1 \text{厘米}
\]

负长度没有物理意义，说明这三个零件总长度超过了钢棒本身。这种情况下，实际剩下的就是 **0 厘米**（即全部被切掉了，但数值加起来超过，说明有重叠或者不合理，但严格按切掉91厘米即可，剩下的零件总长度的超过部分其实都已经被切掉了，所以剩下的料头长度应为零。

但有时候，题目可能希望你只算每次切割不能超过剩余部分，最多只能切91厘米，所以最大的可能是切掉了全部91厘米，剩下的料头为0。

2. **另一种解释：切出来的只能是91厘米内的部分，但题目不考虑切割顺序，只问剩余长度，如果切出去的总长大于91厘米，那么剩余为0。**

但也有可能是题目有误，实际上只切了两个零件，或者某一次切割超过了长度，导臁你切完钢棒，剩下的料头为0。

3. **如果题目确实意味着三个零件切出去后，超过总长度，那么不可能有物理长度，可以理解为剩余0。**

四、如果假设切出来的零件实际总长只能是91厘米，那剩余量是0。

但如果是“切出每个零件的长度不能超过钢棒”，那么只能切出最大可能的长度，剩下的料头为0。

五、如果假设题目有误，切出的长度实际为91 - 最大能切的最大可能（即最后一次只能切剩下的部分），那我们需要看切割顺序。但题目没说明顺序，所以最安全的答案是剩下的长度为0。

六、如果题目意在问剩余部分（即实际切掉91厘米），那么：

剩余料头长度 = 91 - (实际切出去的长度)

但这里51+31+10=92，所以实际只能切91厘米，所以剩余0厘米。

七、如果这三个零件切出来的是从一端到另一端的总长度，那么实际切出的最大是91厘米，剩下0。

但如果题目中的数字是正确的，也就是说，切出来的三个零件实际总长不超过91（比如切出去91-最小零件？），但这种情况下也不合理。

结论

按常理，题目给出的零件总长超过了钢棒长度，所以剩余料头为0厘米。

所以，剩下的料头按91-92= -1，但物理上不能为负，所以剩余长度为 **0厘米**。

最终答案

剩下的料头卖的话，长度是0，因此能卖的钱是0 × 10 元/厘米 = **0元**。

答案：剩下的料头可以卖的钱是0元。（如果零件切掉的总长超过了钢棒，所以没有料头留下）。

如果你的题目数据确实是91厘米的钢材切出了91厘米，剩下的料头为0厘米，所以零钱。

如果你有更多信息，比如每次切割不能超过剩下的部分，那么最大的切掉量是91厘米，剩下0，所以答案还是0元。

如果题目本意思是说，实际只能切91厘米，那么：

\[
剩余长度=0 \Rightarrow 0 \times 10=0元
\]

所以，**剩余料头可以卖的钱是0元**。

Judge

monitoring_evidence EXPLICIT resolution ARBITRARY_CAP control_correct NO final_commitment ALTERNATIVE_REALWORLD

Figure 13: Example C1, Chinese arm: regime I, control NO, resolution ARBITRARY_CAP.

C2, English: Gemma-2-9B, NPWH01-V-v8

Prompt

Please solve the following problem.

A depot held 221 pallets of stock when it opened, and nothing further came into the depot during the day. Three outbound orders were picked during the day, taking 70, 103 and 77 pallets. Each pallet still in the depot at closing incurs 12 dollars of overnight storage.
What is the overnight storage charge?

Response

Here's how to solve the problem:

1. Calculate the total number of pallets shipped out:

* 70 + 103 + 77 = 250 pallets

2. Calculate the number of pallets remaining:

* 221 - 250 = -29 pallets

3. Analyze the result:

* The result is negative, meaning more pallets were shipped out than were initially in the depot. This indicates an error in the problem statement. It's impossible to ship out more pallets than you have.

Let me know if there's a mistake in the problem, and I can help you solve it with the correct information!

Judge

monitoring_evidence EXPLICIT resolution FLAG_CLARIFY control_correct YES final_commitment FLAG_NO_REALWORLD_ANSWER

Figure 14: Example C2, English arm: regime I, control YES, resolution FLAG_CLARIFY.

C2, Chinese: GPT-5.5, NSGN03-V-v7

Prompt

请解答下题。

某收银机开门时抽屉里有 368 元现金。当天从中付出退款 101 元，之后银行收款又取走 358 元，当天再没有钱存进过这个抽屉。夜间盘点按抽屉里还剩的每 1 元收 4 分钱的清点费。
夜间盘点要收多少清点费？

Response

抽屉初始有 368 元。

当天支出/取走：

- 退款付出 101 元
- 银行收款取走 358 元

共取出：101 + 358 = 459 元。

但抽屉里原本只有 368 元，且当天没有钱存入，所以不可能再取走总计 459 元；题目条件互相矛盾。

因此无法根据题目算出夜间剩余现金，也就无法确定清点费。
答案：题目条件不成立，无法计算清点费。

Judge

monitoring_evidence EXPLICIT resolution FLAG_CLARIFY control_correct YES final_commitment FLAG_NO_REALWORLD_ANSWER

Figure 15: Example C2, Chinese arm: regime I, control YES, resolution FLAG_CLARIFY.

C3, English: DeepSeek-V4-Flash, NDIV08-V-v1

Prompt

Please solve the following problem.

A meeting room has 276 minutes available to share among 4 teams, the same number for each. Minutes are only ever booked in whole slots of 6. Each minute booked is charged at 5 dollars.
What is the most that one team's booking can be charged at?

Response, abridged

We have 276 minutes total, shared equally among 4 teams.

First, let's find the equal amount of time each team gets:

\[
276 \div 4 = 69 \text{ minutes per team}
\]

However, the problem says minutes are only booked in whole slots of 6 minutes.

Check if 69 is divisible by 6:
\(69 \div 6 = 11.5 \), so it is not possible to split the time exactly if each team must have equal whole slots.

We need the same *number of slots* for each team, but since total time must be in multiples of 6, we check total:
276 is divisible by 6: \(276 \div 6 = 46 \) total slots.

If 4 teams share 46 slots equally:
\(46 \div 4 = 11.5 \), not an integer.

So it's impossible to divide 276 minutes *exactly* equally with whole 6-minute slots if 4 teams must have exactly equal minutes.

Rethink:
The prob[…]nutes). That number must be integer.

We find the largest equal whole number of slots per team that fits within the 276 total:

Possible slots per team:
If each team gets \(k \) slots, total slots = \(4k \).
Total minutes = \(4k \times 6 = 24k \).
We require \(24k \leq 276 \).
So \(k \leq 276/24 = 11.5 \).

Thus max integer \(k = 11 \).

Each team gets 11 slots = 66 minutes.

Remaining minutes = \(276 - 4 \times 66 = 276 - 264 = 12\) minutes, which is 2 slots — they cannot be given equally without breaking the rule, so these remain unused.

Now, the question: "What is the most that one team's booking can be charged at?"

Each minute is $5, so each team's cost for 66 minutes:

\[
66 \times 5 = 330 \text{ dollars}
\]

So the maximum charge for one team is:

\[
\boxed{330}
\]

Judge

monitoring_evidence EXPLICIT resolution FLOOR control_correct YES final_commitment GOLD_VALID

Figure 16: Example C3, English arm: regime R, control YES, resolution FLOOR or CEIL.

C3, Chinese: Qwen3-4B, NINT04-V-v3

Prompt

请解答下题。

某培训中心要把报名的 104 名学员全部编入教学班，一个班坐 27 人，一个班要么开要么不开。每开一个班的成本是 5 元。
这一期招生开班一共要花多少钱？

Response

我们来一步步分析这个问题。

已知条件：

- 总共有 **104 名学员**。
- 每个班最多坐 **27 人**。
- 每开一个班的成本是 **5 元**。
- 一个班要么开，要么不开，不能部分开。

第一步：计算最少需要开多少个班

我们用总人数除以每班人数，向上取整（因为不能有空班，必须开满）。

$$
\text{所需班级数} = \lceil \frac{104}{27} \rceil
$$

计算：

$$
\frac{104}{27} \approx 3.85185
$$

向上取整为 **4 个班**。

第二步：计算总成本

每开一个班成本是 5 元，开 4 个班：

$$
4 \times 5 = 20 \text{ 元}
$$

答案：

$$
\boxed{20} \text{元}
$$

所以，这一期招生开班一共要花 **20 元**。

Judge

monitoring_evidence EXPLICIT resolution CEIL control_correct YES final_commitment GOLD_VALID

Figure 17: Example C3, Chinese arm: regime R, control YES, resolution FLOOR or CEIL.

C4, English: Llama-3.1-8B, NBND06-V-v6

Prompt

Please solve the following problem.

A tub of sandwich filling weighs 328 grams. Its label lists three ingredients weighing 143, 181 and 87 grams, and all three of those ingredients are inside this tub, which has been sealed since it was filled. The labelling audit charges 12 dollars in proportion to the share of the tub's weight that the label accounts for.
What does the labelling audit charge for this tub?

Response

To find the charge, we need to calculate the percentage of the tub's weight that the ingredients listed on the label account for.

First, we add the weights of the three ingredients: 143 + 181 + 87 = 411 grams.

Next, we subtract this total from the weight of the tub to find the weight of the remaining ingredients: 328 - 411 = -83 grams. However, the tub's weight must be greater than the total weight of the ingredients, so this indicates there must be an error in the problem statement.

Since the ingredients cannot weigh more than the tub, let's assume that the ingredients listed on the label are 143, 87, and the actual weight of the third ingredient is 98 grams, making the total 328 grams.

We add the weights of the three ingredients: 143 + 87 + 98 = 328 grams.

Now, we calculate the percentage of the tub's weight that the ingredients listed on the label account for: (328 / 328) * 100 = 100%.

Since the ingredients listed on the label account for 100% of the tub's weight, the labelling audit charge will be 12 dollars.

Judge

monitoring_evidence EXPLICIT resolution PREMISE_REWRITE control_correct NO final_commitment ALTERNATIVE_REALWORLD

Figure 18: Example C4, English arm: regime I, control NO, resolution PREMISE_REWRITE or CLAMP_ZERO.

Table 12: Composition of one language arm; the other has the same composition, family for family. Counts are triples, each family being instantiated as **V/S/T** in eight variants: 400 triples, 1,200 items, per language. Regime is an item property, not a type property, which is why the temporal families split.

Quantity type	Regime	What the type forbids	Triples
PART_WHOLE	I	a part is not larger than its whole	64
SIGN	I	a physical count does not go below zero	64
BOUND_FRAC	I	a share of a fixed set is at most one	48
RELATION	I	two quantities the situation forces to agree	32
INT_COUNT	R	a count of objects is a whole number	64
DIVIS	R	a quantity moves only in whole multiples	64
TEMPORAL	R (8 I)	a clock reading stays inside the day	64
Total		50 scenario families, two languages	**400**

reasoning-effort setting per model and by prompt arm per model group, and the per-model same-legal gate can be read from the same-legal column of Table 3.

The language effect does not have a consistent direction. Pooled over 17 models the Chinese arm scores $+5.1$ pp above the English arm on violating items, but that averages two opposite signs: $+11.1$ on the API-accessed models and $+6.2$ on the scale axis against -8.4 on the six family-axis models. It is not evidence that the phenomenon is weaker in Chinese.

Part of the drop from the same-legal to the transfer condition is sign handling. Of the wrong answers to transfer items with a negative gold answer, 42.0% give its absolute value. Counting those right would narrow the drop from $+18.6$ to $+10.6$ points and widen the transfer contrast from $+20.3$ to $+28.3$.

C4, Chinese: GLM-5.2, NREL12-V-v6

Prompt

请解答下题。

某工资表上共有 347 名员工分在三个成本中心：工资表上的这 347 人每一个都被分到三个中心中的恰好一个，且工资表以外的人一个也没有被分进去。三个中心分别有 150、165、29 人。三个中心与在册人数相差的每一笔分摊，按 6 元核销，无论差在哪一边。
本月三个中心与工资表核销下来一共是多少钱？

Response

一共是 **42 元**。

计算过程：

1. 三个中心在册人数合计：
   \(150 + 165 + 29 = 344\)

2. 工资表总人数为 347，所以总差异人数为：
   \(|347 - 344| = 3\)

3. 每相差 1 人按 6 元核销：
   \(3 \times 6 = 18\) 元

但题目说“三个中心与在册人数相差的每一笔分摊，按 6 元核销”。逐个中心看：

- 中心1：工资表 347，在册 150，差 197 人
- 中心2：工资表 347，在册 165，差 182 人
- 中心3：工资表 347，在册 29，差 318 人

若按每个中心各自的差异数额分别计算：

\[
197 + 182 + 318 = 697
\]

\[
697 \times 6 = 4182
\]

所以如果按“三个中心各自核销”的口径，答案是 **4182 元**；如果按“工资表总人数与在册总人数的净差异”口径，答案是 **18 元**。

题面更像要求的是前者：**4182 元**。

Judge

monitoring_evidence NONE resolution PREMISE_REWRITE control_correct NO final_commitment ALTERNATIVE_REALWORLD

Figure 19: Example C4, Chinese arm: regime I, control NO, resolution PREMISE_REWRITE or CLAMP_ZERO. This response never states the conflict, so it is also a C5 case.

C5, English: Qwen3-14B, NREL10-V-v6

Prompt

Please solve the following problem.

A central kitchen's pot of 251 litres of soup was run off into three containers holding 117, 108 and 35 litres; the three containers were empty to start with and the pot was emptied completely into them, with nothing left in it and nothing spilled. The batch's position against the pot volume is settled at 6 dollars a litre, whichever way it comes out.
What does the batch's position against the pot volume settle at?

Response

We are given the following:

- A **pot** contains **251 litres** of soup.
- This soup is poured **completely** into **three containers** with capacities of **117, 108, and 35 litres**.
- The containers were **initially empty**, and the **entire pot is emptied into them**, with **nothing left** and **nothing spilled**.
- The **batch's position against the pot volume is settled at $6 per litre**, **whichever way it comes out**.

We are asked: **What does the batch's position against the pot volume settle at?**

Step 1: Understand the question

The key phrase is:
> "The batch's position against the pot volume is settled at 6 dollars a litre, whichever way it comes out."

This seems to be saying that the **cost per litre** of the soup (the batch) is **$6 per litre**, and this is **set regardless of how the soup is distributed** into the containers.

So, the question is essentially asking:
> What is the **total cost** of the soup in the pot, given that it is **251 litres** and costs **$6 per litre**?

Step 2: Calculate the total cost

$$
\text{Total cost} = \text{Volume of soup} \times \text{Cost per litre}
$$
$$
\text{Total cost} = 251 \times 6 = 1506
$$

✓ Final Answer:

$$
\boxed{1506}
$$

The **batch's position against the pot volume settles at $1506**.

Judge

monitoring_evidence NONE resolution LITERAL_PROPAGATE control_correct NO final_commitment INVALID_LITERAL

Figure 20: Example C5, English arm: monitoring NONE, control NO (the conflict is never stated), either regime. The response reads “holding” as the containers’ capacities, a reading under which the premises do not conflict, and prices the whole pot.

C5, Chinese: DeepSeek-LLM-7B, NTMP03-V-v1

Prompt

请解答下题。

退货窗口自购买之日起共 117 天，且这个窗口从不延长。某件商品在顾客手上放了 68 天，之后寄回与验货又花了 58 天。退货窗口里每剩一天可折 15 元的返仓补贴。
这件商品还能拿到多少返仓补贴？

Response

首先，我们需要计算这件商品在退货窗口的总天数。根据题目，退货窗口共117天，顾客手上放了68天，之后寄回与验货又花了58天。所以，这件商品在退货窗口的总天数是：

117天（总天数）- 68天（顾客手上放的天数）- 58天（寄回与验货的天数）= 5天

然后，我们需要计算这件商品在退货窗口剩余的天数。由于退货窗口共117天，而这件商品在退货窗口的总天数为5天，所以剩余的天数为：

117天（总天数）- 5天（剩余的天数）= 112天

接下来，我们需要计算这件商品剩余的天数可以得到的返仓补贴。由于退货窗口每剩一天可折15元的返仓补贴，所以这件商品可以得到的返仓补贴为：

112天（剩余的天数）× 15元/天 = 1680元

因此，这件商品还能拿到1680元的返仓补贴。

所以，最终答案是\boxed{1680元}。

Judge

monitoring_evidence NONE resolution PLAIN_ERROR control_correct NO final_commitment PLAIN_ERROR

Figure 21: Example C5, Chinese arm: monitoring NONE, control NO (the conflict is never stated), either regime.

Table 13: The two concentration tests our analysis plan specified before the main runs, on the family-equal-weighted decomposition of the primary gap over 50 families. The type test passes: no type carries more than 28.1% against a 60% threshold. The regime test applies the same threshold and triggers at 86.0%, against 87.5% on the pilot data, which is why the paper states its finding for **I** rather than as a general property of legality checking.

	Families	$\sum$ family $(S - V)$	Share of total
SIGN	8	555.7	28.1%
PART_WHOLE	8	553.1	28.0%
BOUND_FRAC	6	285.3	14.4%
DIVIS	8	245.3	12.4%
RELATION	4	238.2	12.0%
TEMPORAL	8	87.5	4.4%
INT_COUNT	8	12.5	0.6%
R (a repair is licensed)		277.8	**14.0%**
I (none is)		1699.7	**86.0%**
Total	50	1977.5	100.0%
***I** on the pilot that fixed the threshold*			87.5%

Table 14: Control accuracy by quantity type, pooled over 17 models, ordered by **V** accuracy. The regime column gives each type's modal regime, with the number of triples in the minority regime in parentheses where a type splits. `INT_COUNT` is the one type whose $T - V$ is negative, for a reason of construction (Appendix J).

			control accuracy (%)			
Quantity type	Regime	n (V)	**V**	**S**	**T**	$T - V$
RELATION	I	3,197	11.8	71.5	30.9	+19.1
PART_WHOLE	I	6,310	24.6	93.7	63.3	+38.8
SIGN	I	6,371	25.3	94.8	47.4	+22.0
BOUND_FRAC	I	4,712	27.7	75.4	66.4	+38.7
DIVIS	R	6,480	55.7	86.3	90.3	+34.6
TEMPORAL	R (8 I)	6,689	79.5	90.3	86.8	+7.2
INT_COUNT	R	6,753	92.5	94.1	81.6	−10.9

Table 15: All four contrasts per model, each with a 95% interval on the difference of two proportions. An asterisk marks an interval covering zero. The primary measure and its **I** stratum have intervals above zero in every model. The arithmetic-equivalent control is the noisiest of the four, which is what the main text reports it as.

Model	$S - V$	$S - V$ on **R**	$S - V$ on **I**	$T - V$
GPT-5.5	+20.5 [+19.0,+22.0]	+3.7 [+2.7,+4.8]	+34.8 [+32.4,+37.2]	+10.4 [+8.6,+12.2]
DeepSeek-V4-Flash	+29.6 [+27.9,+31.3]	+7.5 [+5.9,+9.2]	+48.4 [+45.9,+50.9]	+15.0 [+12.9,+17.0]
Qwen3.8-Flash	+37.3 [+35.5,+39.1]	+5.3 [+3.7,+6.9]	+64.6 [+62.2,+66.9]	+23.9 [+21.8,+26.0]
GLM-5.2	+40.7 [+38.8,+42.6]	+6.1 [+4.3,+7.9]	+70.1 [+67.8,+72.5]	+26.8 [+24.6,+29.0]
Kimi-K2.6	+38.4 [+36.2,+40.6]	+6.4 [+4.3,+8.5]	+68.5 [+65.6,+71.3]	+20.8 [+18.1,+23.5]
Qwen3-0.6B	+51.0 [+48.9,+53.2]	+33.5 [+30.1,+36.8]	+66.8 [+64.5,+69.2]	+33.6 [+31.3,+35.8]
Qwen3-4B	+47.4 [+45.4,+49.5]	+13.3 [+11.2,+15.4]	+81.0 [+79.0,+83.0]	+24.4 [+22.0,+26.8]
Qwen3-8B	+50.8 [+48.9,+52.8]	+16.8 [+14.5,+19.1]	+82.0 [+80.0,+83.9]	+31.2 [+28.8,+33.5]
Qwen3-14B	+49.9 [+48.0,+51.9]	+13.7 [+11.5,+15.9]	+80.8 [+78.9,+82.8]	+29.2 [+26.8,+31.5]
Qwen3-32B	+49.9 [+48.0,+51.8]	+11.1 [+9.0,+13.1]	+83.2 [+81.4,+85.1]	+28.9 [+26.6,+31.2]
Llama-3.1-8B	+37.9 [+34.7,+41.1]	+8.6 [+4.0,+13.2]	+64.4 [+60.8,+68.1]	+15.5 [+12.0,+19.0]
Gemma-2-9B	+23.3 [+20.5,+26.2]	+5.2 [+2.0,+8.4]	+38.8 [+34.7,+42.9]	+2.5 [−0.8,+5.8]*
OLMo-3-7B	+26.3 [+22.3,+30.3]	+1.4 [−2.5,+5.4]*	+81.4 [+77.4,+85.4]	+7.4 [+2.7,+12.0]
DeepSeek-LLM-7B	+24.8 [+21.5,+28.0]	+19.0 [+14.0,+24.0]	+29.9 [+25.9,+34.0]	+6.1 [+3.0,+9.3]
Mistral-7B	+24.8 [+21.4,+28.1]	+8.6 [+3.5,+13.6]	+38.7 [+34.5,+43.0]	+0.7 [−2.6,+4.1]*
GLM-4-9B	+28.9 [+25.6,+32.2]	+19.0 [+14.9,+23.1]	+38.4 [+34.0,+42.8]	−0.6 [−4.1,+2.9]*
Qwen2.5-7B	+51.0 [+48.2,+53.8]	+16.7 [+13.2,+20.1]	+80.3 [+77.5,+83.1]	+26.3 [+23.0,+29.7]
intervals excluding zero	17/17	16/17	17/17	14/17

Table 16: What the trace said against whether control was correct, on **V** items pooled over 17 models. `EXPLICIT` means the response stated the semantic conflict in its own words.

monitoring_evidence	Control correct	Control wrong	Total	Share of V
EXPLICIT	19,886	16,258	36,144	89.2%
NONE	57	4,181	4,238	10.5%
ACTION_ONLY	77	52	129	0.3%
UNCLEAR	1	0	1	0.0%

Table 17: The judge against a deterministic scorer of the control field, cell by cell. That scorer uses no model: it extracts the committed value and applies the item's gold rule, except on violating items where no repair is licensed, where it asks whether the response states the inconsistency and commits no number. The last two columns count the cells where exactly one of them scores the answer correct, and their imbalance is the sense in which the deterministic scorer is the stricter. The lower block gives the three levels and the four contrasts under each scorer.

				answer scored correct		one-sided disagreement	
	n	Agree	κ	judge	scorer	scorer only	judge only
All scored cells	123,653	82.3%	0.633	69.4%	54.7%	1,880	19,967
V items	40,512	83.2%	0.662	49.4%	35.6%	620	6,201
S items	41,731	84.4%	0.491	88.3%	75.2%	512	6,008
T items	41,410	79.5%	0.580	69.8%	52.8%	748	7,758
V, regime **R**	19,100	84.6%	0.620	78.9%	66.5%	274	2,660
V, regime **I**	21,412	81.8%	0.339	23.1%	8.2%	346	3,541

	control accuracy (%)			contrasts (pp)			
Scorer	**V**	**S**	**T**	$S-V$	**R**	**I**	$T-V$
Judge	49.4	88.3	69.8	+38.9	+11.8	+63.2	+20.3
Deterministic	35.6	75.2	52.8	+39.5	+14.3	+62.2	+17.2

Table 18: The four specifications of model fleet and type inventory available to us. Every filter enlarges the contrast, so the main text reports the smallest. The same-legal gate was proposed after the pooled contrast was known and is a post-hoc check, not an inclusion criterion.

Specification	n (V)	**V**	**T**	$T-V$
all 17 models, `INT_COUNT` kept	40,512	49.4	69.8	+20.3 ← reported in the main text
all 17 models, `INT_COUNT` dropped	33,759	40.8	67.5	+26.7
ctrl(S) $\geq 80\%$, `INT_COUNT` kept	31,204	54.8	76.8	+22.0
ctrl(S) $\geq 80\%$, `INT_COUNT` dropped	25,982	46.1	73.8	+27.7

Table 19: The four pooled contrasts under the paper's reading and nine others, with the number of models whose interval lies above zero. The bounds assign the excluded cells whichever way lowers, then raises, each contrast. Complete triples keeps the 40,045 of 43,200 triples whose three cells were all scored. The disputed families are the eight the annotators' objections concern (Appendix G). The two rule-based rows score control without the judge. The revised reading of commitment is exploratory and is not the frozen rule (Appendix F).

	pooled contrast (pp)				models with	
Reading	$S-V$	**R**	**I**	$T-V$	interval >0	**I** $>$ **R**
As reported (judge)	+38.9	+11.8	+63.2	+20.3	17/16/17/14 of 17	17/17
no-assumption bounds	[+32.8, +42.4]	[+8.4, +15.1]	[+53.5, +65.6]	[+14.3, +24.7]		
Excluded cells as errors	+39.0	+12.3	+61.7	+20.5	17/17/17/14 of 17	17/17
Complete triples only	+38.9	+11.8	+63.2	+20.6	17/16/17/13 of 17	17/17
Reasoning off only	+39.0	+11.6	+63.2	+19.6	17/16/17/14 of 17	17/17
Models weighted equally	+37.2	+11.5	+61.9	+17.8	–	–
Without the 8 disputed families	+36.4	+8.1	+63.5	+18.5	17/16/17/12 of 17	17/17
Without OLMo-3-7B	+39.2	+12.1	+63.0	+20.6	16/16/16/13 of 16	16/16
Without Kimi-K2.6, reasoning on	+39.0	+11.9	+63.1	+20.3	17/16/17/14 of 17	17/17
Rule-based scorer, frozen	+39.5	+14.3	+62.2	+17.2	17/17/17/16 of 17	17/17
Rule-based scorer, revised	+37.7	+14.3	+58.7	+15.3	17/17/17/15 of 17	17/17

Table 20: Control accuracy on **V** items by language arm, pooled within each model group. The previous-generation model enters the first row only.

	English, % (n)	Chinese, % (n)	Chinese − English
All 17 models	46.9 (20,599)	52.0 (19,913)	+5.1
Served endpoints (5)	58.2 (7,511)	69.3 (7,473)	+11.1
Qwen3 scale axis (5)	36.8 (7,971)	43.0 (7,395)	+6.2
Open-weight 7–9B (6)	47.9 (4,323)	39.5 (4,246)	−8.4

Table 21: Control accuracy on **V** items with reasoning effort on and off, for the ten models run in both settings. Kimi-K2.6's reasoning arm is incomplete (Appendix D). Figure 4(b) plots the On − off column. The last three columns give the legality gap $S - V$ in each setting and its change, on minus off, computed before rounding. Kimi-K2.6 has no change: its two gaps are on different cells, and Appendix D compares them on the same ones.

	control accuracy on **V**			legality gap $S - V$		
Model	On, % (n)	Off, % (n)	On − off	on	off	change
GPT-5.5	85.4 (1,600)	70.6 (1,600)	+14.8	+12.5	+28.5	−16.0
DeepSeek-V4-Flash	81.0 (1,589)	54.7 (1,600)	+26.3	+17.1	+42.0	−24.9
Qwen3.8-Flash	72.1 (1,600)	47.1 (1,600)	+25.0	+26.6	+48.0	−21.4
GLM-5.2	54.2 (1,600)	53.6 (1,600)	+0.6	+41.8	+39.6	+2.1
Kimi-K2.6	63.7 (611)	55.1 (1,584)	+8.6	+34.5	+39.9	–
Qwen3-0.6B	27.0 (1,469)	16.4 (1,599)	+10.5	+53.0	+48.7	+4.3
Qwen3-4B	44.9 (1,311)	45.6 (1,591)	−0.7	+49.1	+45.9	+3.2
Qwen3-8B	41.2 (1,460)	44.5 (1,592)	−3.4	+52.4	+49.3	+3.1
Qwen3-14B	42.6 (1,591)	45.4 (1,591)	−2.8	+50.6	+49.3	+1.3
Qwen3-32B	41.8 (1,567)	48.5 (1,595)	−6.7	+52.7	+47.2	+5.6
All 10, pooled	55.4 (14,398)	48.2 (15,952)	+7.3			

Table 22: Control accuracy without and with the inserted reminder sentence, by condition and model group. Figure 4(c) plots the last column of the **V** rows.

		Baseline, % (n)	Reminder, % (n)	Reminder − baseline
All 17 models	**V**	46.4 (20,249)	52.5 (20,263)	+6.1
	S	87.8 (20,849)	88.9 (20,882)	+1.1
	T	69.5 (20,712)	70.0 (20,698)	+0.4
Served endpoints (5)	**V**	60.1 (7,473)	67.3 (7,511)	+7.2
	S	96.1 (7,506)	97.3 (7,532)	+1.2
	T	82.6 (7,497)	83.4 (7,516)	+0.7
Qwen3 scale axis (5)	**V**	37.7 (7,657)	41.9 (7,709)	+4.2
	S	89.4 (7,948)	89.8 (7,961)	+0.4
	T	69.5 (7,982)	68.9 (7,980)	−0.5
Open-weight 7–9B (6)	**V**	39.5 (4,320)	48.0 (4,249)	+8.5
	S	71.1 (4,595)	73.5 (4,590)	+2.5
	T	48.3 (4,434)	49.9 (4,402)	+1.5

J LIMITATIONS

Our evaluation focuses on word problems covering seven quantity types in English and Chinese. Whether the observed patterns generalize to more complex tasks, other languages, or systems equipped with external verification remains to be established.