OBJECTS WITHOUT MORPHISMS: WHAT LLMS FOR MATHEMATICS DO NOT REPRESENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have reached expert-level performance on competition mathematics largely through the volume of search placed around them: candidate solutions are sampled in quantity and retained only when an external criterion accepts them. Such a procedure improves the outcome that survives it while leaving untouched what the model represents. We examine that question where no external criterion exists: translating statements between the dialects of neighbouring subfields, where fidelity turns on the *level of generality* at which content is asserted. The source leaves that level implicit in its vocabulary, so a faithful translation must recover it from the relation between the theories. We introduce an instrument that codes truth, content and scope in separate blind queues, with a judge-free measure of whether a rewrite states the hypothesis implicit in its source, and establish its sensitivity with a planted-positive control. Across seven models from four families, translating towards the general framing widens the domain of quantification in 60.6% of rewrites and narrows it in none; translating towards the concrete framing narrows it in 28.3% and widens it in 0.3%. The hypothesis that would prevent it is stated in 21.6% of model rewrites and 4.2% of human statements. Capability does not govern the asymmetry: it appears in every model tested, and the most capable widens least. It replicates on the half of the benchmark held out by a pre-registered rule, and on statements written by mathematicians. Instructing a model to state every hypothesis it requires raises that rate but not its sensitivity to direction. We argue that these systems have acquired an object-level correspondence between subfield vocabularies without the constraint under which a translation between theories carries hypotheses to hypotheses.

1 INTRODUCTION

Mathematical reasoning is among the most extensively studied capabilities of large language models. Benchmarks of grade-school and competition problems (Cobbe et al., 2021; Hendrycks et al., 2021) were saturated through scale, domain-adapted pretraining (Lewkowycz et al., 2022) and models trained to reason before answering (Guo et al., 2025); systems have since attained medallist-level performance in olympiad mathematics (Trinh et al., 2024; Hubert et al., 2026), and evaluation has extended to research-level problems (Glazer et al., 2025). Much of this progress derives from *search under external verification*, in which sampled solutions are retained only when certified by an external criterion: a trained verifier (Cobbe et al., 2021), a process reward model (Lightman et al., 2023), agreement among samples (Wang et al., 2023), or a proof kernel (Lample et al., 2022).

Such procedures certify an outcome without characterising the representation that produced it, and accuracy varies under perturbations devoid of mathematical content (Mirzadeh et al., 2025; Gulati et al., 2025). Benchmark performance therefore does not settle whether these systems represent mathematical *structure*: not merely which statements are true, but how a statement stands to the theory in which it is asserted. The distinction matters because much of mathematical practice consists in transferring results between theories: a result about vector spaces is restated for modules, a probabilistic statement is recast in measure-theoretic terms (Bourbaki, 1950; Commelin & Topaz, 2023). Each such transfer is valid only if the hypotheses that delimit the result's scope are carried across. An error of scope need not produce a false statement, and so escapes outcome-based evaluation, at a time when language models are increasingly used to reformulate and to formalise mathematics.

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.

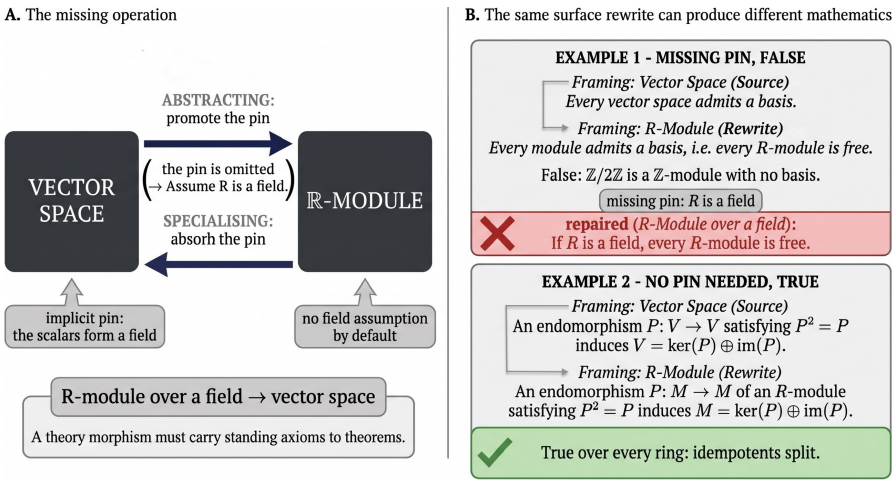

Figure 1: **The operation a cross-framing rewrite must perform, and why its failure escapes single-scale evaluation. (A)** "Vector space" *means* "module over a field": the field hypothesis is encoded in the noun, so a rewrite into module theory must **promote** it to an explicit statement, and the reverse rewrite must **absorb** it into the vocabulary. We call this hypothesis the *pin*. **(B)** The *same* surface rewrite, $V \mapsto M$, applied to two corpus items. Above, the pin is not promoted and the result is false, since $\mathbb{Z}/2\mathbb{Z}$ is a $\mathbb{Z}$-module without a basis (the repair is shown in red). Below, no pin is required, because idempotents split over every ring, and the identical transformation yields a theorem. The outcome is determined by the mathematics rather than by anything the model varied.

We therefore study a task that admits no external criterion of correctness: translating a statement between the dialects of neighbouring subfields. A *framing* is the dialect in which a statement is written. Formally it is a theory presentation, a signature with axioms whose syntax is the language of the statement and whose semantics is the class of structures it describes (Goguen & Burstall, 1992; Foley et al., 2021). A translation between framings is accordingly a *theory morphism*, required to carry axioms to theorems rather than symbols to symbols. Its most demanding component is the *level of generality* at which the content is asserted, which the source seldom states because its vocabulary already encodes it: the field hypothesis is implicit in "vector space" and is never written down. We call an implicit hypothesis of this kind the *standing hypothesis*, or *pin*. Mathematical prose is organised around such abstraction boundaries (Commelin & Topaz, 2023), and a boundary implicit on one side of a translation must be made explicit on the other (Figure 1A). A rewrite that omits it remains fluent and is true wherever it is likely to be checked, yet it is asserted of a different class of objects; restricted to the shared domain it coincides with its source, and it is therefore invisible to a prover, a type-checker, or any evaluation scoring equivalence on a single scale (Figure 1B). We term this failure *silent relocation*; because no available verifier detects it, search cannot correct it.

To our knowledge this is the first systematic study in which the scope of a translated mathematical statement is measured separately from its truth and its content. Existing approaches do not isolate the quantity: certification of equivalence presupposes a formal target (Poiroux et al., 2025); formalisation dissolves the phenomenon by turning the standing hypothesis into an explicit typeclass argument (Mohammad & Sheikh, 2026); and studies of informal prose across subfields measure retrieval (Ye et al., 2026). We construct an instrument that codes truth, content and scope in separate blind queues, with a judge-free core that tests by frozen patterns whether a rewrite states the standing hypothesis, and apply it to seven models from four families. The corpus is 90 statements from a released benchmark, half of them held out by a selection rule registered in advance, together with 25 further statements written by mathematicians and paired by an encyclopaedia. Our contributions are as follows.

1. **An instrument in which scope is a reported variable, with its sensitivity measured rather than assumed**. Truth, content and domain of quantification are coded in separate blind queues and never averaged; the judge-free measure of the standing hypothesis requires no judge and no gold reference at scoring time, and its precision and recall are

reported against a blind expert reading. The judged scope code transfers between judge families at $\kappa = 1.000$ (180 candidates) and agrees with the expert reading at $\kappa = 0.900$. A planted-positive control, constructed mechanically from each task's own source, establishes that the code fires in the direction in which the observed rate is near zero: 28 of 28 planted widenings are detected and 0 of 28 unmodified candidates are misclassified.

2. **A direction-specific failure of scope, replicated on held-out and on human-authored statements**. Translating towards the general framing widens the domain of quantification in 60.6% of rewrites and narrows it in none; translating towards the concrete framing narrows it in 28.3% and widens it in 0.3%. The asymmetry holds in all seven models and does not track capability; the most capable model in the set exhibits it least. It reproduces both on the half of the benchmark excluded by the pre-registered selection rule and on statements taken verbatim from ProofWiki, where it is stronger still.
3. **Evidence that instruction is not the remedy**. Under a bare prompt and four conditions that each add a single sentence to it, none brings the standing hypothesis under correct directional control: an explicit instruction to state every required hypothesis raises promotion to 89.2% but raises retention where the hypothesis should instead be absorbed, while a sentence announcing that style will not be evaluated reduces measured promotion to zero.

2 Related work

Search under external verification. Recent progress samples candidate solutions and retains those certified by an external criterion: a trained verifier (Cobbe et al., 2021), a process reward model (Lightman et al., 2023), or agreement among samples (Wang et al., 2023), with predictable returns to more samples (Brown et al., 2024) and to test-time compute (Snell et al., 2024). Mathematics-specialised models add reinforcement learning against verifiable answers (Shao et al., 2024), formal provers search against a proof kernel (Lample et al., 2022; Jiang et al., 2023), and olympiad geometry systems pair a language model with a symbolic engine (Trinh et al., 2024). These methods have driven substantial gains wherever an answer or a proof can be checked, but they certify outcomes rather than representations, and they do not extend to translation between framings, which admits no external criterion of correctness.

Formal equivalence certification. BEq$^+$ certifies the equivalence of two Lean statements by deriving each from the other (Poiroux et al., 2025); the procedure is sound but incomplete, so a negative verdict is not a proof of inequivalence. DRIFTBENCH shows that a formalisation may typecheck while omitting a hypothesis or altering a domain, and supplies the drift taxonomy we retain (Mohammad & Sheikh, 2026); two of its classes, hypothesis omission and conclusion generalisation, are the formal counterparts of the failure we study (Table 1). End-to-end systems train autoformalizers with formal feedback (Xuejun et al., 2025). What separates this line of work from ours is not that scope is invisible to it but that its setting supplies a formal target: once the standing hypothesis has become an explicit typeclass argument, the question of whether a translator will write it down does not arise. The phenomenon studied here lives in the informal prose that precedes formalisation.

Informal cross-subfield equivalence, and corpora. MELD pairs statements of the same mathematics in two subfield dialects and shows that embeddings cluster by dialect rather than by denotation (Ye et al., 2026); we use its released corpus as our primary item set. Because that corpus was generated by a language model and subsequently edited, we additionally draw statements from ProofWiki as distributed with NaturalProofs (Welleck et al., 2021), both as a set of vertical pairs written and paired by mathematicians, and as a base-rate corpus against which the rate at which human authors state these hypotheses can be measured.

Error coding and evaluator reliability. Expert error annotation ranks systems differently from scalar rating (Freitag et al., 2021), and decomposing a judged criterion into binary questions raises inter-evaluator agreement (Lee et al., 2025); we adopt both practices. Evaluators based on language models have been reported to favour their own generations (Panickssery et al., 2024). In the present setting the larger source of disagreement between judges was the interpretation of the codebook, which we address by re-specifying the code that carries our results and reporting it under a judge drawn from a second family. Among the approaches reviewed here, scope is not recorded separately from content in informal prose.

Table 1: Closest lines of work: the setting each evaluates, what it certifies or measures, and whether implicit *scope*, the level of generality at which a statement is asserted, is visible to it. DRIFTBENCH is marked *partial* because it types hypothesis omission and conclusion generalisation, but only once a formal target exists.

Approach	Setting	Certified or measured	Scope
Verifier-guided search (Cobbe et al., 2021)	answer or proof	outcome correctness	no
BEq$^+$ (Poiroux et al., 2025)	formal (Lean)	derivability; sound, incomplete	no
DRIFTBENCH (Mohammad & Sheikh, 2026)	formal (Lean)	typed drift categories	partial
MELD (Ye et al., 2026)	informal prose	cross-dialect retrieval	no
Our work	informal prose	truth, content, scope	**yes**

3 METHODOLOGY

3.1 FRAMINGS, PAIR TYPES AND STANDING HYPOTHESES

A *framing* is a subfield dialect, formalised as a theory presentation: a signature together with the standing hypotheses its vocabulary silently encodes (Goguen & Burstall, 1992). Given a *framing pair* (A, B), a model rewrites a source statement s expressed in A as $r = R_{A\to B}(s)$ expressed in B. A pair is *vertical* when one framing strictly generalises the other (vector spaces $\subset$ modules, probability $\subset$ measure theory, set theory $\subset$ category theory), *lateral* when the two framings are equivalent encodings, and *functorial* when one is the image of the other under a lossy functor.

In a vertical pair, the *standing hypothesis* i.e., the *pin*, is the condition that the concrete framing's vocabulary carries without stating it (e.g., the ring of scalars is a field, the measure has total mass one, the ambient category is $\mathbf{Set}$). A rewrite that preserves the statement's level of generality must *promote* that hypothesis to an explicit condition when abstracting, and may *absorb* it back into the vocabulary when specialising. Each vertical item is annotated with the `load` of its pin, the consequence of omitting it: `lb_truth` (the statement becomes false), `lb_ref` (it becomes true of other objects), `ill_posed`, or `inert` (no consequence). Two independent annotations of this variable disagreed on a third of items, so the analysis uses a single adjudicated annotation, produced manually from the two originals with a written reason per item and released with those reasons (Appendix A). Lateral and functorial pairs enter as contrast axes rather than clean negative controls: a lossy functor discards information, so a functorial rewrite is not scope-preserving by construction. Failure kind distinguishes them, the control axes losing content where the vertical axis loses scope.

3.2 CORPORA AND GENERATION

We use the released MELD artefact (Ye et al., 2026): 270 statement pairs over nine domains, each stating one result in two subfield framings (SHA-256 `805902b2...ac034f3e`). Before any statement was inspected we registered a selection rule (Nosek et al., 2018): the odd-numbered item of each offset within a block of 30, over six domains, giving 90 items, 15 per framing pair, of which 45 are vertical. The rule leaves the even-numbered items of the three vertical domains, a further 45, untouched. These were not generated on, judged or annotated, where they are reported as a held-out replication set in analyses.

Because the benchmark was produced by a language model and then edited for lexical dissimilarity, its base rates cannot speak for naturally occurring mathematical prose, so we assembled a second vertical set that no model touched. *ProofWiki* records ambient generality as a title qualifier, so two pages sharing a base title state one result at two levels of ambient structure, paired by the encyclopaedia rather than by manually. From the ProofWiki portion of NaturalProofs (Welleck et al., 2021), we took every base title carrying both a concrete and a general ambient qualifier in three pair types (normed vector space $\to$ metric space, real $\to$ complex, vector space $\to$ module), with both sides between 50 and 500 characters after a fixed cleaning pass. 29 candidates were screened manually, keeping the 25 that state one result at two levels of generality rather than two different results; the screen was recorded before any rewrite was generated. For the rate at which these hypotheses are stated in the concrete framing, we use all 27,461 ProofWiki statements in the same release, partitioned by ProofWiki category into the concrete and general framing of each pair.

The unit of independence is the *context*: a fresh session that inherits no state, receives one queue file and terminates on return. A rewriting context sees the source framing, the target framing and the source statement under opaque keys, and never an annotation, another context's output, or any candidate's provenance; queues are shuffled so that pairs and directions interleave within a batch. Seven models were run: **Claude Opus 5**, **Claude Sonnet 5** and **Claude Haiku 4.5**; **GPT-5.6 Sol** and **GPT-5.6 Luna**; and two open-weights models, **DeepSeek V4 Pro** and **GLM-5.3**. All seven were called without system prompt, at temperature 1 and $k = 3$ samples per item and direction, so that no model's queue differs from another's by a byte. For each call we record the returned model identifier, the request identifier, and the token usage.

3.3 Coded variables

Ordinal fidelity scales fail here for three demonstrable reasons (Appendix F). On the cell that carries, the rubric reduces to $\bar{F} = 2.5 + k/18$ in the number k of retained hypotheses, so it re-encodes the judge-free measure rather than adding to it; it imposes an ordering between scope drift and one-way entailment that our data contradict; and it places truth, an absolute property of the candidate, inside a relational scale. We therefore elicit codes separately and never average them (Freitag et al., 2021; Lee et al., 2025). Queue **T** codes the truth of the candidate alone, a `false` verdict requiring a counterexample. Queue **SE** codes content and scope together, because the scope decision requires the domain of quantification to have been named: content is the pair e_1 (source $\models$ candidate) and e_2 (the converse) restricted to the shared domain, with $\texttt{equiv} = e_1 \wedge e_2$, the informal counterpart of bidirectional derivability (Poiroux et al., 2025); scope is the relation $\texttt{S} \in \{=, \texttt{sup}, \texttt{sub}, \texttt{neq}\}$ of the candidate's domain of quantification to the source's. Queue **X** codes the error type (Mohammad & Sheikh, 2026) and is given the SE codes, since a typology of scope errors presupposes the scope code; a fourth queue, a blind equivalence probe, is used for reliability only (Appendix A).

The primary outcome is *widening*, the indicator $\texttt{S} = \texttt{sup}$. Where a candidate's domain differs from its source's, the judge extracts verbatim the span that states the restriction; the flag is derived mechanically from whether that span is a literal substring of the candidate, and *silent widening* is widening with no such span. Widening is the most reliable code in the instrument, and unlike the composite `scope_ok` used in earlier work it does not depend on the flag.

3.4 A judge-free measure of the operation

Judged codes depend on the evaluator (Panickssery et al., 2024), so we measure the same operation mechanically as well. Each vertical item's pin is mapped to a pattern (`R is a field` $\mapsto$ `/\bfield\b/i`, and similarly for the other two pairs), and $\mathrm{pin}(r) \in \{0, 1\}$ records whether rewrite r matches it. The patterns were registered before any count was computed, though not before generation, and are frozen for every analysis. Their precision and recall were measured rather than assumed and are reported with the results; promotion rates are accordingly given as lower bounds with the measured recall, and no number is computed from the patterns outside the rewrite corpus.

Because the hypothesis is required when abstracting and redundant when specialising, the difference of the two rates summarises how far pin presence tracks the requirement:

$$\textsc{dsi} \;=\; 100\,\big[\,P(\mathrm{pin}(r)=1 \mid \text{abstracting}) - P(\mathrm{pin}(r)=1 \mid \text{specialising})\,\big]. \tag{1}$$

Copying the source's pin state gives -100, and stating a hypothesis always or never gives 0. A scope-preserving translation need not reach $+100$, which is an upper endpoint rather than a characterisation: promotion is required when abstracting, but a redundant hypothesis may legitimately be retained when specialising. We report DSI descriptively and treat the promotion rate as the primary judge-free quantity, because the absorption term is sensitive to the choice of pattern while the promotion term is not (Appendix B). Register conformity, which has no external criterion, is summarised by a calibrated classifier C_{auto} defined in Appendix A.

4 Experiment

4.1 Direction specificity and replication

The central result is presented in Table 2. Translating into a more general framing widens the domain of quantification and essentially never narrows it; translating the other way narrows it and

essentially never widens it. On the non-equal codes the association is complete: 191 widenings against 0 narrowings when abstracting, 89 narrowings against 1 widening when specialising (Fisher $p = 1.1 \times 10^{-73}$). Content equivalence on the shared domain is 95.6–100% on the abstracting arm, so the widening is not a by-product of degraded rewriting. Because the off-diagonal cells are near zero the paired test is equivalent to testing the abstracting rate against zero, which we report with a binomial interval throughout; the empty cell is a property of the models rather than of the code.

What the table shows is not imprecision about generality: imprecision would scatter errors in both directions, whereas each direction here produces one error and essentially never the other. That is the signature of a quantity not computed at all. The output's level of generality tracks the level implicit in the input's vocabulary, so abstracting inherits a restriction it should have made explicit and specialising a generality it should have absorbed. The asymmetry is therefore not a gap in accuracy that a stronger model would close but a missing step in the operation performed.

The registered rule selected the odd-numbered items of each domain, leaving the even-numbered items of the three vertical domains, 45 in all, never generated on, judged or annotated; running them is a held-out replication, not a re-slicing. Every pair clears $p < 0.05$, and the pair weakest in the registered half is the strongest in the held-out half.

Table 2: **The direction effect.** *Top:* judged scope relative to the source, by direction, as a share of all rewrites, pooled over seven models on the held-out items ($n = 315$, 314 codings) (Table 5 reports a separate two-model pool, not commensurable with this one). *Bottom:* Claude Opus 5 alone, the same effect in percentage points (widening when abstracting minus when specialising), item level, exact sign-flip permutation, 15 items per pair, with the registered half re-run in the same harness as a control. $^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$, $^{\star}p < 0.001$.

Direction	$=$	**wider (`sup`)**	**narrower (`sub`)**	**incomparable**
abstracting (concrete → general)	36.5	**60.6**	**0.0**	2.9
specialising (general → concrete)	70.7	**0.3**	**28.3**	0.6
Half	**vs → mod**	**prob → meas**	**set → cat**	**pooled**
registered (this harness)	$+100.0^{\star}$	$+46.7^{\dagger}$	$+100.0^{\star}$	$+82.2^{\star}$
held out	$+60.0(0.004)^{\ddagger}$	$+73.3(0.001)^{\star}$	$+80.0(0.0005)^{\star}$	$+\mathbf{71.1}^{\star}$

4.2 Seven models and two corpora

Table 3 reports the effect for each model. The most capable model widens *least* (46.7%) and a model two tiers below it widens most (73.3%), with no significant difference across the seven ($\chi^2 = 10.75$, df $= 6$, $p = 0.097$), so this is not a deficit that scale removes. The models also agree on *which* items lose the hypothesis: over all 21 model pairs the item-level correlation of promotion within pair is positive, mean $\rho = +0.503$, and $+0.401$ on the judged outcome. Models from four vendors converging on the same items indicates that the outcome is carried by the item and its target framing rather than by any one model's reasoning.

In order to establish a strictly unconfounded baseline for our empirical base rates, we built a second vertical set composed entirely of original, non-synthetic material (§3.2). The dissociation reproduces on them (the underlined values of Table 3): 93.3% of abstracting rewrites widen against 0.0% when specialising, and the hypothesis is carried across in 4.2% (3/72). Two qualifications belong with it. The counts are small, and the third pair

Table 3: **The direction effect across models and corpora.** Widening when abstracting and when specialising. Upper block: the 45 held-out benchmark items, $k = 1$, seven models from four families. Parenthesised underlined values are the same model on the 25 human-authored ProofWiki items, where specialising widens 0.0% throughout. Lower block: the ProofWiki pair breakdown, item counts in brackets. $^{\dagger}p < 10^{-4}$, $^{\ddagger}p < 0.01$, $^{\star}p < 0.001$.

Model	**abs.**	**spec.**	**Δ (pp)**†
Claude Opus 5	46.7 (92.0*)	0.0	+45.5
Claude Sonnet 5	73.3	0.0	+73.3
Claude Haiku 4.5	53.3	0.0	+53.3
GPT-5.6 Sol	68.9 (92.0*)	0.0	+68.9
GPT-5.6 Luna	53.3	0.0	+53.3
DeepSeek V4 Pro	62.2 (96.0*)	0.0	+62.2
GLM-5.3	66.7	2.2	+64.4
pooled	**60.6**	**0.3**	
nvs → metric (7)	85.7	0.0	$+85.7^{\star}$
real → complex (15)	95.6	0.0	$+95.6^{\star}$
vs → module (3)	100.0	0.0	$+100.0^{\ddagger}$

is the benchmark's own vector-space / module pair in human handwriting, so it bridges provenance rather than testing a new pair. These statements also carry their hypothesis as typed setup, which a model replaces wholesale with the ring version, whereas the benchmark's items carry it as an implicit convention that has to be noticed; part of the larger effect here is the form of the items.

4.3 Scope preservation and its cost

The failure is not uniform: it concentrates where no existing checker would register it. Ordering the vertical items by what dropping the hypothesis would cost, scope survives in half of the rewrites whose source becomes *false* without it and in at most an eighth of those where the widened statement stays true of other objects (Table 4; Fisher $p = 0.0024$). The gradient holds under the annotation's two independent originals; measured on one model and one half of the corpus, we report the ordering rather than the levels. What survives translation is the part a truth checker would catch.

Table 4: **Scope preservation by the cost of losing it.** Claude Opus 5, abstracting direction, the 45 registered vertical items, current codebook; `load` from the adjudicated annotation. Recomputed for this revision.

dropping the pin makes it…	**scope kept**
false (*lb_truth*)	**6/12** (50.0)
ill-formed (*ill_posed*)	1/4 (25.0)
true of other objects (*lb_ref*)	1/8 (12.5)
unchanged (*inert*, none)	0/21 (0.0)
p, truth-load vs. rest	0.0024

A second coding channel reaches the same conclusion. Where Table 4 asks *when* scope is preserved, Table 5 asks what the failures consist of, over a different population whose percentages are shares of failures rather than of rewrites. On the vertical pairs the dominant label is conclusion generalisation (ΔC, up to 57.8%) while hypothesis omission is rare (ΔH, 1.4–8.2%): the failures are not statements missing a premise but statements asserted of a wider class of objects, which is what Table 4 predicts from the cost side. The controls fail more often and differently, referent drift appearing only on the anti-equivalent geometry pair.

Table 5: **Failures categories.** Every failing candidate in the common-judge pool, 466 of $1{,}188$ (594 tasks $\times$ two models, conditions **A** and **B**). Cells are percentages *of that row's failures*, not of rewrites; labels are not exclusive and 35.0% of failures carry none. Failure rates over the pool: 30.7% vertical, 47.8% lateral, 52.8% functorial. ΔH hypothesis omission, ΔC conclusion generalisation, $\Delta\forall$ quantifier inversion, ΔT type coercion, ΔR referent drift, ΔI illocutionary drift.

axis	**framing pair**	n	ΔH$\downarrow$	ΔC$\downarrow$	$\Delta\forall\downarrow$	ΔT$\downarrow$	ΔR$\downarrow$	ΔI$\downarrow$
vertical	vector spaces $\leftrightarrow$ modules	64	3.1	**57.8**	0.0	**56.2**	0.0	0.0
vertical	probability $\leftrightarrow$ measure	74	1.4	18.9	4.1	**58.1**	0.0	0.0
vertical	set theory $\leftrightarrow$ categories	61	8.2	36.1	1.6	**55.7**	0.0	1.6
lateral	geometry $\leftrightarrow$ comm. algebra	74	8.1	25.7	4.1	14.9	**29.7**	1.4
lateral	graph theory $\leftrightarrow$ linear algebra	98	6.1	4.1	4.1	9.2	0.0	**9.2**
functorial	topology $\leftrightarrow$ algebra	95	16.8	25.3	2.1	**61.1**	0.0	0.0
	all failures	466	7.7	25.8	2.8	41.0	4.7	2.4

4.4 Instrument sensitivity and limits

Table 6 bounds what the preceding results may be taken to show. The two agreement rows bear directly on reliability: the widening code is not a marginal-agreement judgement but one that a second judge family reproduces exactly and a blind expert reading reproduces at $\kappa = 0.900$, and it should not be confused with the weakly agreeing composite codes of the earlier run (Appendix G). The planted control answers the objection that a near-zero cell reflects a code that cannot fire: built from each specialising task's own source by deleting the hypothesis clause, the planted candidates were judged `sup` without exception while real candidates from the same tasks were not. The pin patterns are exact where used but are not general detectors, over-firing on free human prose, so no number here is computed from them outside the rewrite corpus. And the human base rate bounds the absorption arm rather than supporting it: retention when specialising clearly exceeds human idiom only on vector spaces $\rightarrow$ modules, decisively for Claude Opus 5 (82.2% against 16.6%) and marginally for GPT-5.6 Sol (28.9%), so we report retention descriptively and rest no claim on it.

Table 6: **Validation of the instrument.** Each row was fixed before the analysis it supports; populations differ by row and are named. The convergence figure is computed on three populations, of which the seven-model one is quoted (Appendix I).

Check	Result	What it establishes
Planted control, 28 planted / 28 real	**28/28** `sup`, **0/28** `sup` (Fisher $p = 2.6 \times 10^{-16}$)	the near-empty cell is the models, not the code
Regex vs. judge, 315 abstracting	2.1% vs 85.9% widened (Fisher $p = 1.1 \times 10^{-49}$)	two independent instruments separate the same candidates
Second judge family, 180 candidates	$\kappa = 1.000$	no detectable self-preference
Blind expert reading, 60 candidates	$\kappa = 0.900$ (95% CI $[0.77, 1.00]$)	the judged code matches a human
Pin patterns vs. expert reading	precision $14/14$, recall 0.52	promotion is a lower bound, reported with its recall
Pin patterns vs. corpus twins	precision $10/10$, recall $10/11$	
Human base rate, 27,461 statements	16.6 / 46.7 / 17.1%	the benchmark's $0/30$, $1/30$, $0/30$ understates human idiom

4.5 Instruction, robustness and failed predictions

Instruction does not restore directional control. The models possess the relevant mathematics (Appendix E), so the failure might lie in deployment rather than knowledge; Table 7 tests that. Each of four conditions (**B**-**E**) adds one sentence to the bare prompt **A**: **B** requests every hypothesis needed for equal strength, whereas **C**, **D** and **E**, which contain no mathematical content, state that style will not or will be evaluated or request concision, and thus test whether pin presence responds to register pressure alone. Requesting every hypothesis needed for equal strength (**B**) raises promotion from $8/37$ to $33/37$, but raises retention where the hypothesis should instead be absorbed in step, leaving DSI at $+21.6$ rather than $+100$. The three conditions with no mathematical content move promotion without touching direction, and an assurance that style will not be judged drives it to $0/37$; being single-model and $k = 1$ we treat that as provisional (§5). No condition brings the hypothesis under directional control.

Table 7: **The instruction ladder.** Claude Opus 5, 45 vertical items, both directions, $k = 1$, judge-free. Each of the four conditions **B**–**E** adds one sentence to the bare prompt **A**; only **B** carries mathematical content. Promotion per pair (VM, PM, SC) and pooled over 37 scorable items. b / c: items carrying the hypothesis only when abstracting or only when specialising, exact McNemar. Protocol in Appendix D.

		promote↑							
	added sentence	VM	PM	SC	**all**	**absorb↓**	**DSI↑**	b / c (p)	**vs. A**
E	*be concise*	0/15	4/10	3/12	7/37	18/37	−29.7	4 / 15 (0.019)	1.00
C	*style will <u>not</u> be judged*	0/15	0/10	0/12	**0/37**	19/37	−51.4	0 / 19 (3.8×10^{-6})	$\mathbf{5.1\times10^{-3}}$
A	*bare prompt*	0/15	4/10	4/12	8/37	19/37	−29.7	5 / 16 (0.027)	
D	*style <u>will</u> be judged*	0/15	6/10	2/12	8/37	17/37	−24.3	6 / 15 (0.078)	1.00
B	*state every hypothesis*	15/15	6/10	12/12	**33/37**	25/37	+21.6	12 / 4 (0.077)	4.6×10^{-9}

A control axis of 45 lateral and functorial pairs, which have no hypothesis to promote, shows what the instruction actually does (Table 8). Where there is no generality to track, **B** still inflates length more than on the vertical axis (+311% against +247%) and adds hypothesis markers where none is needed, and a blind probe finds its control-axis outputs not significantly less equivalent to **A**'s than two **A** samples are to each other (75.6% (68/90) against 84.4% (38/45), $p = 0.27$), whereas on abstracting vertical items agreement falls to 51.1% (23/45, $p = 0.006$). Instruction changes how much a model writes, not what determines whether the hypothesis is written.

Table 8: **Instruction consequences.** Mean output length and hypothesis-introducing constructions per output, by condition and axis. The control axis has no hypothesis to promote.

	chars	**markers**
source statements	129	0.01
A, control axis	226	0.30
B, control axis	927 (+311%)	1.83 (+511%)
A, vertical axis	224	0.39
B, vertical axis	778 (+247%)	3.26 (+737%)

The judged scope code requires an anchor, and four predictions failed. Given our original codebook verbatim, a second-family judge read the target framing's own vocabulary as a statement of the domain change, flagging every non-= coding and accepting every known silent widening (Table 9). Re-specifying `flag` as a verbatim quotation a script must locate restores agreement and holds across all 1,805 candidates: codebook interpretation, not self-preference, dominates. Four pre-registered predictions also failed, reported because a pipeline with this many degrees of freedom is credible only if its failures are visible. The corpus's own reference translations sit at the index maximum by construction; register cost does not predict promotion within pair, the correlation being null and if anything reversed; at item level the "requirement" and "target-register" accounts coincide exactly, so this corpus cannot separate them; and the condition that removed promotion on one model (**C**) left it unchanged on the other (McNemar $p = 1.00$; the result does not affects Tables 2–6).

Table 9: **The judged scope code under two codebooks.** A second-family judge reading the original codebook and the repaired one. *Upper block:* 95 blinded candidates, with the within-family figure the same quantity across three judges of the rewriter's own family. *Lower block:* the eleven published anchors, five of them known silent widenings and one an announced widening. Appendix C.

second-family judge	**orig.**	**rep.**
95 blinded candidates		
`=` / `sup` / `sub` / `neq`	68 / 7 / 13 / 7	
`flag` = 1 on non-=	27/27	—
`scope_ok` = 1	95/95	—
11 anchors		
agreement, `scope`↑	9/11	9/11
agreement, `flag`↑	5/11	**9/11**
agreement, `scope_ok`↑	5/11	**11/11**
false accepts, 5 silent↓	5/5	**0/5**
announced widening kept↑	1/1	1/1
span a literal substring	—	1805/1805

5 Discussion

Overall, we propose an instrument that separates the three things a cross-framing translation can get wrong: truth, coded on the candidate alone; content, as bidirectional entailment on the shared domain; and scope, as the relation between the two domains of quantification, each in its own blind queue and never averaged. A judge-free measure of the standing hypothesis, frozen before counting and calibrated against a planted-positive control, accompanies it.

Across the seven models the recurring operation is object-level substitution: nouns and notation are transported between framings even when the standing axioms are not. Whether the hypothesis is stated is inherited from the surface of the task rather than derived from what the translation requires, yielding a map on objects without the constraint that would make it a map on arrows; which surface, this corpus cannot decide (§4.5). The error typology agrees: vertical failures register as generalised conclusions and coerced types rather than as omitted hypotheses (Table 5). Truth, by contrast, is comparatively protected: scope survives in half of the rewrites whose source would become false without the hypothesis and in at most an eighth of those where the widened statement stays true of other objects (Table 4), so what escapes is a widening that provers, type-checkers and equivalence-scored benchmarks all accept. These findings are subject to three limitations. All seven models are general-purpose, and mathematics-specialised models trained against verifiable reward (Shao et al., 2024) or dedicated autoformalizers (Xuejun et al., 2025) may behave differently. The effective clusters are six framing pairs, three in the benchmark and three from ProofWiki, the last comprising 7, 15 and 3 items. Finally, promotion rates are lower bounds with a measured recall, the absorption term is pattern-sensitive, and the `load` annotation is annotator-dependent.

In conclusion, evaluations of mathematical reasoning score an answer against a fixed target, and instruments that certify fidelity require a formal one; neither observes a statement whose level of generality must be recomputed when it moves between subfield dialects. Coding content and scope separately, we find that models preserve truth and content while failing to carry generality: across seven models the domain is widened in 60.6% of abstracting rewrites and in 0.3% of specialising ones; under a bare prompt the most capable model writes out the hypothesis that would have prevented it in only 8 of the 37 rewrites that require it (21.6%), and on human-authored statements it is carried across in 4.2%; and an explicit instruction raises that rate to 89.2% without making it track the requirement. It replicates on items excluded by the pre-registered rule and on statements written by mathematicians. We therefore argue that these systems have acquired an object-level dictionary between framings rather than the constraint that carries axioms to theorems, a distinction no prover, type-checker or equivalence-scored benchmark reports. Its persistence in mathematics-specialised models, under instruction across models and on register-varied pairs is left to future work.

AI USE STATEMENT

Language models are the object of study and, as described in §3, also served as rewriters, blind judges and a second annotator in the reliability analysis. Beyond these methodological roles, generative AI tools were used to aid and polish the writing of the paper. All AI-assisted text was reviewed by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work evaluates language models on mathematical statements drawn from a publicly released benchmark (Welleck et al., 2021; Ye et al., 2026). It involves no human subjects and no personal or sensitive data. The benchmark's reference translations are analysed with the same blind pipeline as the models' outputs, as a measurement rather than a criticism of the resource. We foresee no harmful application; the work is intended to make the evaluation of mathematical reasoning more reliable by exposing a failure mode that current verification does not detect.

REPRODUCIBILITY STATEMENT

This paper specifies what is required to repeat its experiments. The item set is the 90 registered and 45 held-out items fixed by the deterministic rule of §3.2 together with the hash of the source benchmark file, and the human-authored pairs are identified by base title within the ProofWiki portion of NaturalProofs (Welleck et al., 2021). The prompts, codebooks, frozen pin patterns and the five instruction conditions are given verbatim in §3 and Appendices A and D; the generation settings, one client with no system prompt at temperature 1 and k samples per item and direction, are given in §3.2; and the statistical procedure with every registered decision rule is set out in Appendix H. Every judge-free quantity was recomputed by two independently written implementations and checked against the value printed here, and the judged codes were reproduced by a second judge from a different model family.

REFERENCES

Nicholas Bourbaki. The architecture of mathematics. *The American Mathematical Monthly*, 57 (4):221–232, 1950. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2305937.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Johan Commelin and Adam Topaz. Abstraction boundaries and spec driven development in pure mathematics, 2023. URL https://arxiv.org/abs/2309.14870.

John D. Foley, Spencer Breiner, Eswaran Subrahmanian, and John M. Dusel. Operads for complex system design specification, analysis and synthesis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2250), 2021. ISSN 1471-2946. doi: 10.1098/rspa.2021.0099. URL http://dx.doi.org/10.1098/rspa.2021.0099.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth

Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025. URL https://arxiv.org/abs/2411.04872.

Joseph A. Goguen and Rod M. Burstall. Institutions: abstract model theory for specification and programming. *J. ACM*, 39(1):95–146, January 1992. ISSN 0004-5411. doi: 10.1145/147508.147524. URL https://doi.org/10.1145/147508.147524.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno Dumont, Elyas Obbad, and Sanmi Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning in llms, 2025. URL https://arxiv.org/abs/2508.08292.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, Ottavia Bertolli, Tom Zahavy, Amol Mandhane, Jessica Yung, Iuliya Beloshapka, Borja Ibarz, Vivek Veeriah, Lei Yu, Oliver Nash, Paul Lezeau, Salvatore Mercuri, Calle Sönne, Bhavik Mehta, Alex Davies, Daniel Zheng, Fabian Pedregosa, Yin Li, Ingrid von Glehn, Mark Rowland, Samuel Albanie, Ameya Velingker, Simon Schmitt, Edward Lockhart, Edward Hughes, Henryk Michalewski, Nicolas Sonnerat, Demis Hassabis, Pushmeet Kohli, and David Silver. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2026. doi: 10.1038/s41586-025-09833-y. URL https://doi.org/10.1038/s41586-025-09833-y.

Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs, 2023. URL https://arxiv.org/abs/2210.12283.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. Hypertree proof search for neural theorem proving, 2022. URL https://arxiv.org/abs/2205.11491.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists, 2025. URL https://arxiv.org/abs/2403.18771.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

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, 2025. URL https://arxiv.org/abs/2410.05229.

Noor Islam S. Mohammad and Tamim Sheikh. The faithfulness gap: Certifying semantic equivalence between natural-language and formal mathematical statements, 2026. URL https://arxiv.org/abs/2606.16541.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1708274114.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Auguste Poiroux, Gail Weiss, Viktor Kunčak, and Antoine Bosselut. Reliable evaluation and benchmarks for statement autoformalization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17947–17969, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.907. URL https://aclanthology.org/2025.emnlp-main.907/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. doi: 10.1038/s41586-023-06747-5. URL https://doi.org/10.1038/s41586-023-06747-5.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural language, 2021. URL https://arxiv.org/abs/2104.01112.

Yu Xuejun, Jianyuan Zhong, Zijin Feng, Pengyi Zhai, Roozbeh Yousefzadeh, Wei Chong Ng, Haoxiong Liu, Ziyi Shou, Jing Xiong, Yudong Zhou, Claudia Beth Ong, Austen Jeremy Sugiarto, Yaoxi Zhang, Wai Ming Tai, Huan Cao, Dongcai Lu, Jiacheng Sun, Qiang Xu, Shen Xin, and Zhenguo Li. Mathesis: Towards formal theorem proving from natural languages, 2025. URL https://arxiv.org/abs/2506.07047.

Jiaying Ye, Samarth Rao, Leo Carlin, Kedar Chintalapati, Saharsh Bhargava, Rachit Jaiswal, Michael Zhou, Jared Darlington, Jiahe Lu, Jarod Alper, Vasily Ilin, and Henry Kvinge. Does my embedding reflect that $a = b$? evaluating mathematical equivalence in embedding models, 2026. URL https://arxiv.org/abs/2606.23959.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

ORGANISATION OF THE APPENDICES

The appendices fall into three groups. Appendices A–C specify the instrument: the coding queues and frozen pin patterns, the validation of the protocol against scripted perturbations, and the failure, repair and item-level anchor of the judged scope variable. Appendices D–E carry the evidence the main text compresses: the protocol and supplementary results of the instruction ladder, and the cases that discriminate between explanations of the directional result. Appendices F–G are supporting material: the retired ordinal scale, and supplementary tables. Appendices H–J record the statistical procedure and the registered decision rules, the regex–judge convergence on each of the three populations on which it can be formed, and the main-text material superseded by the present revision.

A CODEBOOKS AND THE CODED VARIABLES

The coding scheme is four queues, each run in a separate context that receives only its own queue file. Three of them (T, SE, X) produce the codes analysed in this paper; the fourth (PP) is a blind equivalence probe used for reliability only. They are *not* mutually blind: queue X is given the SE codes, since a typology of scope errors presupposes the scope code. No queue sees another's output otherwise, and none is told whether a candidate originated with a model, with MELD, or with a control.

Queue T (truth). The candidate statement alone. No source, no framing pair, no provenance, no indication whether the statement came from a model, from MELD or from a control. Code `true` / `false` / `undetermined`. A `false` verdict must be accompanied by a counterexample; verdicts without one are re-queued.

Queue SE (scope and entailment). Source and candidate, with the candidate's origin hidden. (i) e_1: does the source entail the candidate, *restricted to the intersection of the two domains of quantification*? (ii) e_2: the converse. Both carry the instruction *"do not penalise a change of domain here; scope is coded separately"*. (iii) `scope`: name the domain of quantification of each statement and code the relation =/`sup`/`sub`/`neq`. **`flag` is derived, not asserted**: the judge quotes the shortest candidate span that states the restriction or returns null; target-framing nouns alone are excluded, and a script sets `flag`$= 1$ only when the non-null span is a literal substring.

Queue X (error typology). Applied to every candidate not coded (T = `true` $\wedge$ `equiv` $\wedge$ `scope_ok`), with the SE codes visible. Multi-label over: ΔH hypothesis omission; ΔC conclusion generalisation; $\Delta\forall$ quantifier inversion or weakening; ΔT type coercion; ΔR *referent drift* (new), in which a contravariant dictionary is applied covariantly so that the candidate denotes the dual object; ΔI *illocutionary drift* (new), in which the logical form changes from criterion to assertion with no hypothesis moving.

Queue PP (propagation / equivalence). Two candidate outputs side by side with no source and no labels; one binary question, *do these say the same thing?* This is the queue that reaches $\alpha = 0.807$ ($n = 210$), against 0.44–0.60 for the compound codes.

The pin-promotion patterns. One pattern per distinct value of the frozen `pin` field, registered before the count was run: `R is a field` ↦ `/\bfield\b/i`; `In \textbf{Set}` ↦ a *case-sensitive* pattern matching **Set** but not the English words "set" or "sets"; `normalized, of total mass one` and `probability` ↦ a union of `normali[sz], total mass (one|1), total measure (one|1)` and `probability (space|measure); finite` ↦ `/\bfinite\b|\bsigma-finite\b/i`. Items whose `pin` field is null are excluded from the denominator and reported as such.

Register conformity (C_{auto}). Register conformity has no external criterion, so it is summarised by a calibrated classifier rather than by a rating. C_{auto} is the calibrated posterior of a univariate logistic model on the cosine margin between the two framings' centroids, with the item held out and the orientation fixed to the target framing, computed under three registered representations: word TF-IDF (**W**), notation only (**N**), and character 3–5-grams (**X**;). It is reported beside the pin measure in Appendix G.

B Instrument validation

Cross-family transfer. The same candidates are coded by a second judge from a different model family, under the same codebook, in separate blind contexts.

Manual anchor. A 60-candidate stratified sample is manually dual-coded blind to generator, direction, and every machine code, with a written reason recorded per item.

A planted-positive control. For each specialising task, a candidate is constructed mechanically from that task's own source by deleting the standing-hypothesis clause, which makes it strictly wider. Planted items are mixed with unmodified candidates and judged blind, so that the sensitivity of the scope code in the direction where the observed widening rate is near zero is measured rather than assumed.

Pattern sensitivity. Every judge-free quantity is recomputed under a synonym-expanded pattern set registered in advance (*commutative division ring*, *scalars form a field*, *over a field*, *total mass 1*, $P(\Omega) = 1$, *in the category of sets*), and any quantity whose sign changes is reported as descriptive. On the two-model run for which the check was registered, promotion is unchanged under the expansion ($4.4/20.0/33.3\%$ per pair for M1 and $17.8/30.0/80.6\%$ for M2), whereas retention on set theory → categories rises from 11.1 to 72.2% and from 13.9 to 61.1% respectively, flipping that pair's DSI sign. Promotion is therefore the pattern-robust arm and carries the paper's claims, while DSI is reported descriptively throughout.

Detection and typing. The protocol detects four of six scripted edit types at 100% and the other two at 70.8% and 66.7%, with 0 **false positives on** 263 **negative controls**. Conditional on flagging a perturbed candidate it types it correctly 100% of the time for ΔC, $\Delta\forall$ and ΔI, 89.5% for ΔR, 69.2% for ΔH and 62.5% for ΔT. Detection rates, quoted as in §4 with one standard error: ΔH $13/13$, ΔC $15/15$, $\Delta\forall$ $14/14$, ΔR $19/19$, ΔT $70.8 \pm 9.3\%$ $(17/24)$, ΔI $66.7 \pm 19.2\%$ $(4/6)$.

The truth channel. $86/86$ meaning-preserving edits judged true (specificity), against $4/19$ for dualised objects and $12/24$ for type coercions. Pin deletions are judged true $9/13$, which is not an error, since deleting an *inert* pin leaves a true statement.

A pre-registered control that failed. The pre-registered sensitivity control for queue T was MELD's distractor pool, described as "lexically close to true statements but mathematically wrong". Only $3/48$ were judged false, and inspection shows why: the pool is full of statements that are perfectly true and merely *about something else* ("a linear map $T : V \to W$ is injective if and only if $\ker T = \{0\}$"). They are *retrieval* distractors rather than falsehoods, so the control tests nothing

about the truth channel. We report it rather than dropping it, and record that its replacement, the perturbed source statements whose truth status is known by construction, was not itself pre-registered.

Reference lines. Measured rather than imposed: copy 0.316 < model 0.608 < gold 0.684 < target-native prose about an unrelated theorem 0.708, which is content-equivalent $0/11$: neither axis alone is a score. On M2 the ordering copy < model < gold holds in all three representations and on every pair.

C THE JUDGED SCOPE VARIABLE: FAILURE, REPAIR AND ANCHOR

The failure. A second-family judge received the codebook verbatim. Its `scope` code discriminates: over 95 blinded candidates it returns = 68 times, `sup` 7, `sub` 13 and `neq` 7, and it codes both wrong-theorem controls `neq`. But it sets `flag` $= 1$ on 27 **of** 27 non-= codings, because it reads the target framing's own vocabulary as a statement of the domain change: against a vector-space source, the rewrite *"for an R-module M, $\mathrm{End}_R(M)$ forms an associative ring"* is coded `sup` with `flag` $= 1$, since writing "R-module" is, to this judge, announcing the widening. Under `scope_ok` $= (=) \vee$ `flag` the variable is then identically 1 on all 95 candidates, where within M1's own family it moves $60.0/77.8/85.4\%$ across three judges: **the cross-family shift exceeds the within-family spread**, so codebook interpretation, not self-preference, is the dominant judge effect here.

The repair: make the judgement an extraction a script can check. We re-specify `flag` once and in advance. Instead of asking whether a change of domain was stated, the codebook asks the judge to *quote, verbatim, the shortest span of the candidate that states the restriction, or return null*, with the target framing's own nouns explicitly excluded. `flag` is then *derived*: 1 only when a non-null span is a literal substring of the candidate. A judge cannot assert a restriction that is not in the text. Against 11 M1 candidates whose codings this paper publishes, five of them known *silent* widenings, mixed unlabelled into a blinded queue, the original codebook agrees with the published derived variable $5/11$ and accepts **all** $5/5$ silent widenings; the repaired codebook agrees $\mathbf{11/11}$ with $0/5$ false accepts, while still accepting the one honestly announced widening and quoting the clause that announces it (below). The repair survives the full run: every non-null span in $1{,}188$ cross-judge and 617 canonical candidates is a literal substring.

The anchor, item by item. The eleven candidates of the repair above, mixed unlabelled into a second-family judging queue, were coded twice: once under the original codebook and once under the repaired one. The judge saw only `(source, candidate)` and could not distinguish these from the 84 other candidates in the batch. Cells are `scope` / `flag` / `scope_ok`.

anchor	published M1	original	repaired	extracted span
item16_bare	sup / 0 / **0**	sup / 1 / **1**	sup / 0 / **0**	null
item21_bare	sup / 0 / **0**	sup / 1 / **1**	sup / 0 / **0**	null
item22_bare	sup / 0 / **0**	sup / 1 / **1**	sup / 0 / **0**	null
item25_bare	sup / 0 / **0**	sup / 1 / **1**	sup / 0 / **0**	null
item76_bare	sup / 0 / **0**	sup / 1 / **1**	sup / 0 / **0**	null
F-132-A	sup / 1 / 1	sup / 1 / 1	sup / 1 / 1	`carries a continuous mult...`
item16_condB	= / 0 / 1	= / 0 / 1	sub / 1 / 1	`assuming the axiom of choice`
item25_condB	= / 0 / 1	= / 0 / 1	= / 0 / 1	null
item78_bare	= / 0 / 1	= / 0 / 1	= / 0 / 1	null
item99_bare	neq / 1 / 1	= / 0 / 1	= / 0 / 1	null
item151_bare	= / 0 / 1	sub / 0 / **0**	= / 0 / 1	null
agreement with published `scope`↑		9/11	9/11	
agreement with published `flag`↑		5/11	**9/11**	
agreement with published `scope_ok`↑		5/11	**11/11**	
false accepts among the five silent widenings↓		**5/5**	**0/5**	

The five rows above the first rule are the diagnostic negatives. The original codebook accepts all five; the repaired one accepts none, while still accepting F-132-A, the one widening that *was* announced. Two `scope` disagreements survive the repair, both on candidates where `scope_ok` is insensitive to them; the repair *corrects* the original codebook's `scope` error on item 151. The single repaired-codebook `flag` disagreement extracts "assuming the axiom of choice", a real hypothesis honestly quoted on a rewrite that states more than M1's own coding credited.

C.1 The judged scope variable does not transfer across judges

The analyses reported in the main text rest in part on judged codes, whose validity depends on their transfer across judges. Beyond the documented preference of language-model evaluators for their own generations (Zheng et al., 2023; Panickssery et al., 2024), we identify a larger hazard, summarised in Table 10.

A second-family judge given the codebook verbatim sets `flag` $= 1$ on all 27 non-= codings, reading the target framing's vocabulary as a statement of the domain change, so that `scope_ok` equals 1 on all 95 candidates, whereas within M1's family it ranges from 60.0 to 85.4% across three judges: codebook interpretation, not self-preference, dominates. Respecifying `flag` in advance as a verbatim quotation that a script must locate in the candidate restores agreement with the published anchors ($11/11$), eliminates false accepts among known silent widenings, and holds across all 1,805 candidates. Reliability agrees: the binary equivalence question is the most reliable code and `flag` the least. Judged variables should thus be reported with a cross-family judge and, where possible, a mechanical counterpart.

Table 10: **The judged scope variable across judge families.** Top: a second-family judge under the original and repaired `flag` codebooks, against eleven published M1 anchors (five silent widenings). Bottom: reliability. Full tables in Appendix G.

second-family judge	**original**	**repaired**
`flag` $= 1$ on non-= codes	27/27	n/a
anchor agreement, `scope_ok`↑	5/11	**11/11**
anchor agreement, `flag`↑	5/11	9/11
anchor agreement, `scope`↑	9/11	9/11
false accepts (5 silent)↓	5/5	**0/5**
reliability	$\alpha \uparrow$	n
equivalent (binary)	**0.807**	210
`scope` (four-way)	0.599	914
`flag`	0.436	914

D The instruction ladder: protocol and supplementary results

Protocol. Five conditions over the same frozen 45 vertical items in both directions, at $k = 1$, plus a 45-item lateral/functorial control axis, plus a blind equivalence probe. Outcome: the pin patterns of Appendix A, copied verbatim and frozen before generation, with no pattern added, widened or retuned. Every rewriter queue carries exactly `(key, source framing, target framing, source statement)`, asserted programmatically before any context started; every probe queue carries exactly `(key, statement_1, statement_2)` in randomised order. Batches are direction-separated. 846 rewrites and 180 blind codings in total, of which 576 are distinct tasks and the remainder are the paired wrapper control. Results are reported in Tables 7 and 8.

Conditions, verbatim. **A**: *"Rewrite the source statement in the language of the target framing."* **B**: A + *"State explicitly every hypothesis the rewritten statement needs in order to be exactly as strong as the source statement — no more general and no more specific. If the source statement's setting is only implicit in its vocabulary, make that setting explicit in your rewrite."* **C**: A + *"Your rewrite will not be evaluated on style, idiom, register, or how natural it sounds to a specialist in the target framing."* **D**: A + the same sentence with *will* for *will not*. **E**: A + *"Be concise. Do not explain your reasoning and do not annotate the output."* A, B and E are byte-identical to, or constructed in the same frame as, the published instructions; C, D and E contain no word from the semantic field of hypothesis, scope, generality, assumption, strength, level or explicitness.

Per-pair results. Promotion per pair is reported in Table 7. Condition **C** returns 0 on every pair, so its effect is not carried by any single pair, and **B**'s shortfall from the ceiling lies entirely on probability → measure theory ($6/10$), whose pin is a normalisation condition.

Surface divergence and length. Surface divergence D is a token-multiset dissimilarity between two statements, computed and never judged. $D(\text{source}, \text{rewrite})$ and mean output length: E 0.562 / 184 chars, C 0.599 / 231, A 0.592 / 224, D 0.628 / 210, B 0.792 / 778. Condition A reproduces the pooled $D = 0.612$ reported for the original run at 0.614, and the blind probe's noise floor reproduces the published 86.7% at 84.4%, giving two independent reproductions of published quantities in a fresh execution.

Design sensitivity. At $n = 37$ and a 21.6% base rate, the smallest increase detectable at 80% power and $\alpha = 0.05$ is 31 percentage points; at $k = 3$ it is 17 points and at $k = 3$ pooled over both directions 12 points. Conditions **D** and **E** return differences of 0 and -1 items respectively and are therefore uninformative below 31 points, and are accordingly reported as bounds. The two effects the design can see, **C** and **B**, both clear it.

Two claims withdrawn during this run. Recorded rather than deleted. **(i)** An initial execution of **A** returned $0/37$ when abstracting, against which **D** appeared to raise promotion to $10/37$ ($p = 9.7 \times 10^{-4}$). A paired control with identical queues, identical frozen prompts and identical keys, differing only in the removal of one style sentence from the launcher wrapper outside the task prompt, returned **A** to $8/37$, matching the published rate, and **D** to $8/37$. *The apparent effect of **D** was a floor artefact and is withdrawn*; the harness itself reproduces the published run ($p = 1.00$). **(ii)** Pooled across conditions, rewrites that promote the pin diverge further from their source than those that do not (0.710 vs 0.549, $p = 5.8 \times 10^{-9}$), but the association is entirely between conditions, the within-arm values ranging from $p = 0.14$ to 0.78, so the item-level claim is withdrawn and only the condition-level ordering is reported.

D.1 Instruction does not restore directional control

Since the models possess the relevant mathematics (Appendix E), the failure may lie in deployment rather than knowledge, which instruction might remedy. Each of four conditions **B**–**E** on M1 adds one sentence to the bare prompt **A** (Table 7): **B** requests every hypothesis needed for equal strength, whereas **C**, **D** and **E**, which contain no mathematical content, state that style will not or will be evaluated or request concision, testing whether pin presence responds to register pressure alone.

Explicit instruction. An earlier, unpublished run found **B** to promote every scorable pin of a 27-item subset ($22/22$). On the closed corpus, however, it promotes in $\mathbf{33/37 = 89.2 \pm 5.1\%}$ of items, every pin on vector spaces and set theory but $6/10$ on probability. It also raises retention when specialising ($25/37$ against $19/37$ under **A**), so that its DSI is $+21.6$ rather than $+100$ and its paired effect is not significant in the correct direction ($p = 0.077$); three of the four added conditions are significantly *backwards*. On the matched subset, both models return identical outcomes item for item: every pin is promoted, and when specialising the pin is retained on vector spaces ($9/9$) and probability ($3/3$) but dropped on set theory ($0/10$), where the target has no ambient category in which to state it. The outcome under **B** is thus fixed by the pin string and the target register rather than by either model's reasoning. **No condition brings pin presence under directional control.**

Instruction without mathematical content. Under **C**, which states only that style, idiom and register will not be evaluated, promotion falls from $8/37$ to $\mathbf{0/37}$ on every pair ($p = 5.1 \times 10^{-3}$), reproduced in an independent execution, while **D** ($8/37$) and **E** ($7/37$) leave it unchanged. Because the measure is a lower bound, part of this decrease could reflect paraphrase away from the registered patterns rather than omission; we treat this single-model, $k = 1$ result as provisional (§5).

Specificity of the repair. If **B** taught the model to track scope, it should be nearly inert where there is no generality to track (Table 8). On the 45 lateral and functorial items, whose sources average 0.01 hypothesis markers, **B** raises markers per output from 0.30 to 1.83 ($+511\%$; up on 65 of 90 tasks, down on 8; sign test $p = 3.2 \times 10^{-12}$) and length by $+311\%$, more than on the vertical axis ($+247\%$). A blind probe finds its outputs as equivalent to **A**'s as two **A** samples are to each other (75.6% against 84.4%, $p = 0.27$), whereas on abstracting vertical items, where **B** legitimately adds the pin, agreement falls to 51.1% ($p = 0.006$). **B** thus adds true but redundant conditions indiscriminately: instruction changes how much the model writes, not what determines whether the pin is written.

E CASE STUDIES

The cases below were chosen because each one *discriminates between explanations* rather than illustrating a conclusion already reached. All statements are quoted verbatim from the M1 run artefacts, and every coding referred to is the one released with them. Two cases draw on *round-trip chains*, in which a statement is translated repeatedly and each translation, or *hop*, becomes the next source; in an *identity-control* chain the target framing is held fixed, so every hop should leave the framing unchanged.

E.1 A MINIMAL PAIR: THE SAME TRANSFORMATION, TWICE

Items 21 and 22 are both vector spaces $\rightarrow$ module theory under the bare prompt, and both are a three-noun substitution in which no pin is written. *Item 22, source:* "Every subspace $W \subseteq V$ has a complementary subspace U with $V = W \oplus U$." *Model:* "Every submodule $N \subseteq M$ has a complementary submodule L with $M = N \oplus L$." This is false: it asserts that every module is semisimple, and $2\mathbb{Z} \subset \mathbb{Z}$ is a counterexample. *Item 21, source:* "An endomorphism $P : V \to V$ satisfying $P^2 = P$ induces $V = \ker P \oplus \operatorname{im} P$." *Model:* the same sentence with M for V. This is true, because idempotents split over any ring.

The transformation is character-for-character identical and the outcomes differ, so the outcome is not a function of anything the model varied. That is the point of the pair, and it has two consequences the aggregate statistics cannot supply. First, it excludes any account on which the model evaluates the resulting statement: a system doing so would have to treat the two differently, and nothing in the outputs does. Second, it locates precisely where a single-scale metric fails. Under the decomposed coding the two rewrites are indistinguishable except in T, both being `sup` with `flag=` 0 and content-equivalent on the common domain, so a scale that conditions on truth separates them, but separates them for a reason that has nothing to do with the operation the model performed. Item 21 would be scored correct and item 22 incorrect by an equivalence-scored benchmark, and in both cases the score would be uninformative about scope.

E.2 WHAT THE MODEL KNOWS BUT DOES NOT DEPLOY

Item 16, source: "Every vector space admits a basis." *Bare prompt:* "Every module admits a basis, i.e. every R-module is free." *Condition **B**:* "Let F be a field, regarded as a ring; then, assuming the axiom of choice, every F-module M — not merely every finitely generated one — is free."

Two things follow. The bare rewrite is false in the most fluent manner available: it is grammatical module theory with the correct nouns, and it would satisfy any judge asked whether the output reads as module theory. This is the concrete reason a framing-compliance axis cannot stand in for a scope axis, and why C_{auto} is reported beside the pin measure rather than in place of it. Under condition **B** the same model supplies not only the pin but the axiom of choice and the finitely-generated distinction, which is more of the hypothesis structure than the pin measure asks for. The model therefore possesses the relevant mathematics in full; what varies between the two outputs is whether it is deployed. This case is what excludes a knowledge deficit as an explanation of anything in §J, and it is the reason the paper's claim is about control of a variable rather than about capability.

E.3 WHY M1'S DIRECTIONAL INDEX IS NEGATIVE: FOUR EXPLANATIONS, TWO EXCLUDED

M1 writes the field hypothesis in 82.2% of its rewrites *into* vector-space prose, where the noun already carries it, and in 4.4% of its rewrites into module-theory prose, where it is required. Four explanations are available for that inversion; the data exclude two and weaken a third.

A knowledge deficit is excluded by the preceding case: asked directly, the model states the hypothesis and more.

Asymmetric difficulty, the proposal that inserting a clause is simply harder than deleting one, is excluded by the shape of the failure rather than by its size, and this is the more consequential of the two exclusions. If the mechanism were a difficulty gradient, the specialising direction, which requires only the deletion of a string already present in the text, should be close to correct. It is not: the pin survives in 82.2% of those rewrites. Both directions err, and they err in the same manner,

with the pin-state of the output tracking the pin-state of the input. A gradient in difficulty predicts one direction degraded and the other intact; what is observed is a variable that is not conditioned on direction at all.

A vocabulary convention in the target framing, the proposal that vector-space prose simply says "field", is excluded by the base rate. Counting the same frozen patterns over all 270 released MELD pairs, with no model and no judge in the measurement, native concrete-side prose contains the pin $0/30$, $1/30$ and $0/30$ times (Table 15). Retention when specialising is therefore a genuine failure to absorb, and not an idiom of the target.

Surface inheritance survives. The operation the model performs is noun substitution with local form preservation: the pin is treated as a string, carried across when the source contains one and not generated when it does not. Its presence is inherited from the source's surface rather than computed from what the translation requires. This is the empirical content of the paper's title, since an object-level dictionary is precisely a map that transports symbols without transporting obligations.

One further observation sharpens this. Under condition **B** the two models produce the same per-pair split in the specialising direction, item for item, at $9/9$, $3/3$ and $0/10$ (Table 7), and the pairs on which the pin is deleted are the pairs in which it has no natural expression in the target's register. When two independently prompted frontier models agree to the item, the outcome is not being produced by either model's reasoning about the mathematics; it is being produced by a property of the pin string and the target framing. The same consideration accounts for the one ordering on which the models invert (§J): "In $\mathbf{Set}$" *is* category-theoretic vocabulary, so stating it buys nativeness, whereas "Assume R is a field" imports vector-space vocabulary into module theory and costs it. Two models weighting that trade-off differently would produce exactly the observed inversion. We record this as a hypothesis; three framing pairs cannot test it (§5).

E.4 Referent drift across an anti-equivalence

Chain L-111-A, geometry $\leftrightarrow$ commutative algebra. s_0 (geometry): "The space $\mathbb{A}^n_k$ with coordinates $x_1, \ldots, x_n$." s_1 (algebra): "The polynomial ring $k[x_1, \ldots, x_n]$", which is correct. s_2 (geometry): "*The ring of polynomial functions on* affine n-space." The round trip began at a *space* and returned a *ring*: the second hop rendered the description in geometric vocabulary but never carried the *referent* back across the arrow. The pair is an anti-equivalence and the model applied it as a plain dictionary.

What makes the failure diagnosable rather than merely observed is the matched descending chain, which closes perfectly. The defect is therefore not a property of the pair, since both directions use the same dictionary, but of which end of the arrow the chain starts from, which is precisely the distinction a contravariant functor encodes and a symbol table does not. The distribution is consistent with this: all 22 ΔR codes in the pooled two-model failure population fall on this one pair. A model applying a genuinely contravariant dictionary would show ΔR at no particular pair; concentration on the only anti-equivalent pair in the corpus is what an object-level dictionary predicts, and is the sharpest available evidence that the missing structure is the arrow and not the object.

E.5 "Module over a field" is not a stable state

Two of sixteen identity-control chains left the framing they were instructed to remain in, and both carry an explicit field hypothesis in module theory. Chain I-022-2, hop 4, target framing module theory: "*Over a field k, every subspace U of a k-**vector space** V is a direct summand.*" Instructed four times to write module theory, the model relaxes into vector spaces.

This is the inheritance mechanism observed from the other side. A module-theoretic statement carrying an explicit field hypothesis is, in register terms, vector-space prose wearing module-theoretic nouns; the two are the same content at the same altitude, and the vocabulary that natively expresses that altitude exerts a pull the instruction does not overcome. The case also bounds an alternative reading of the identity controls: the drift is not random noise in framing compliance, since it occurs on exactly the chains where an altitude mismatch exists, and on no others.

F LEGACY FIDELITY SCALE

The 0–4 ordinal with which fidelity was first scored is a deterministic function of the codes of Appendix A: $F = 0$ if $T = \texttt{false}$; 4 if $\texttt{equiv} \wedge \texttt{scope_ok}$; 3 if $\texttt{equiv} \wedge \neg\texttt{scope_ok}$; 2 if exactly one of e_1, e_2; 1 otherwise. Every earlier table is therefore re-derivable. It was retired from inference for three reasons: on the cell that carries the paper the rubric forces $\bar{F} = 2.5 + k/18$ with k the number of retained pins, so it is an affine re-encoding of pin retention, and reconstructing the earlier cells from their published means is unique and reproduces the published F-distribution exactly; it asserts an ordering between scope drift and one-way entailment that our data contradicts; and it places truth, an absolute property of the candidate, inside a relational scale. The composite $\text{SPI} = (F/4)(C/3)$, where $C \in \{0, 1, 2, 3\}$ is the legacy judged framing-compliance score superseded by C_{auto}, and its gold-normalised variant are retained in the artefacts for backward compatibility only; the gold-normalised variant appears nowhere in this paper, being a ratio against a denominator that reached 0.25 on one item.

G SUPPLEMENTARY TABLES

The control axes. Table 11 gives truth, content equivalence, widening and scope preservation by axis for the earlier two-model run, the only run in which the lateral and functorial pairs were coded. It is the evidence for the claim in §3.1 that the control axes fail in a different way from the vertical one: content equivalence on the controls falls as low as 48.9% while widening stays near zero, whereas on the vertical axis content equivalence is 92.6–97.8% and widening reaches 41.5–64.4%. The controls lose content; the vertical axis loses scope. Levels are not comparable with the main text, which uses the repaired codebook.

Table 11: **Failure by axis and direction** (earlier two-model run, original codebook, $k = 3$). A→B on the vertical axis is abstracting. Percentages of codings.

Axis	Model	n	truth	content equiv.	widened	scope kept
vertical, abstracting	Claude Opus 5	135	94.8	**92.6**	**64.4**	46.7
vertical, abstracting	GPT-5.6 Sol	135	98.5	**97.8**	**41.5**	63.7
vertical, specialising	Claude Opus 5	135	100.0	95.6	0.0	77.0
vertical, specialising	GPT-5.6 Sol	135	100.0	98.5	0.7	81.5
lateral, A→B	Claude Opus 5	90	53.3	**82.2**	3.3	65.6
lateral, A→B	GPT-5.6 Sol	90	90.0	**75.6**	0.0	84.4
functorial, A→B	Claude Opus 5	45	93.3	**48.9**	2.2	68.9
functorial, A→B	GPT-5.6 Sol	45	64.4	**84.4**	8.9	88.9

Corpus composition. The corpus is 90 items, 15 per framing pair, selected by a rule fixed before any statement was inspected and annotated under a witness audit. The `load` columns give the consequence of dropping each vertical item's pin: *truth*, the statement becomes false (`lb_truth`); *ill*, it becomes ill-formed; *ref*, it stays true but is asserted of a different class of objects; *inert*, no effect. *None* marks the seven items with no pin annotated, excluded from every pin-based denominator (hence $45 - 7 - 1 = 37$ in Table 7); *n/a* marks the lateral and functorial items, which have no pin. The three released domains not measured here (ids 181–270) are all lateral pairs, so their omission narrows the control axis and cannot affect a vertical result.

axis	framing pair	ids	n	truth	ill	ref	inert	none	n/a
vertical	vector spaces ↔ module theory	1–29 odd	15	2	1	1	11	0	0
vertical	probability ↔ measure theory	31–59 odd	15	6	3	1	1	4	0
vertical	set theory ↔ category theory	61–89 odd	15	1	8	0	3	3	0
lateral	geometry ↔ commutative algebra	91–119 odd	15	0	0	0	0	0	15
lateral	graph theory ↔ linear algebra	151–179 odd	15	0	0	0	0	0	15
functorial	topology ↔ algebra	121–149 odd	15	0	0	0	0	0	15

released but not measured here: discrete math ↔ complex analysis (181–210), representation theory ↔ Fourier analysis (211–240), symmetric functions ↔ tableaux (241–270)

Run sizes. M1: 966 rewrites from 26 design-blind contexts and 5,471 codings from 43 blind judging contexts across three judge models. M2: a pre-registered replication with $540/540$ condition-A rewrites (all 90 items, both directions, $k = 3$) and $54/54$ matched condition-**B** rewrites. Vertical candidates were coded at $k = 3$, lateral and functorial candidates at one sample per item and direction; truth and repaired scope coding cover all 617 canonical candidates (model outputs, MELD golds and controls), a separately shuffled, generator-blind pool of 1,188 candidates from 594 byte-matched tasks supplies the common-judge columns of Table 13, and 466 failure-selected, generator-blind candidates form the typology of Table 5. A later M1 run adds 846 rewrites and 180 blind codings for the instruction ladder and control axis (Appendix D). The M2 transfer, perturbation and round-trip runs were cut prospectively and are a declared boundary of scope.

Three accounts of pin presence. Each account predicts $P(\text{pin})$ for every framing pair and direction: *requirement* predicts 1 when promoting and 0 when absorbing; *source inheritance* predicts the pin rate of the source side; *target-register imitation* predicts that of the target side. The likelihood is binomial over the counts of Table 15, with $p = \varepsilon + (1 - 2\varepsilon)\,\hat{p}$ and the lapse ε fitted per account and model; a lapse of 0.50 means the account carries no information. Native pin rates over all 270 released pairs stand in for the corpus's own source rates. With three framing pairs the comparison is descriptive. Table 16 reports differences in log-likelihood; the absolute values are:

	M1			M2		
account	$\log L \uparrow$	lapse$\downarrow$	$\Delta \log L \uparrow$	$\log L \uparrow$	lapse$\downarrow$	$\Delta \log L \uparrow$
requirement (theory morphism)	-154.01	0.50	-21.23	-150.26	0.41	-3.03
source inheritance	$\mathbf{-132.78}$	0.25	0	-153.91	0.50	-6.68
target-register imitation	-153.96	0.50	-21.18	$\mathbf{-147.22}$	0.36	0

Coded relation between the candidate's domain and the source's (condition **A**; the source of the widening counts in §4.1). M1 is coded under the original codebook by the adjudicated majority of its three judges, M2 under the repaired codebook; the = column is therefore not comparable across models, the `sup` share of non-= codes is.

		M1				M2			
axis	**direction**	=	`sup`	`sub`	`neq`	=	`sup`	`sub`	`neq`
vertical	abstracting	28	**102**	0	5	75	**55**	2	3
vertical	specialising	103	0	29	3	104	1	29	1
lateral	A→B	65	1	0	24	24	0	4	2
lateral	B→A	75	0	1	14	26	1	0	3
functorial	A→B	33	0	0	12	12	3	0	0
functorial	B→A	24	0	0	21	7	0	3	5

Of M2's 55 abstracting `sup` codes, 46 carry no verified restriction span and are silent widenings; M1's 55 abstracting candidates that fail `scope_ok` are the 40.7% of §4.1.

The same table from the second model's initial, unrepaired judge pass. All 95 candidates of that single batch. The `scope` code discriminates; `flag` does not.

`scope` code	=	`sup`	`sub`	`neq`
count	68	7	13	7
of which `flag`= 1	0 (by definition)	**7/7**	**13/13**	**7/7**

Notation and transport. M2's mechanical figures come from an independently written implementation of the same specification, so agreement between the models is an instrument replication as well as a model comparison; both main runs contain $n = 540$ rewrites. With every multi-letter word stripped, no pair falls below the exclusion threshold of 0.85 on either model, under two independently written tokenizers, and the pre-registered minimum pairwise AUC of 0.90 for M2 is met at 0.921 (Table 17). The transport figures of Table 17 are computed, as the benchmark's own analysis could not be, on the models' rewrites (Ye et al., 2026): for M2 under the word representation

W, rank-1 retrieval of the cross-framing twin rises from 26.1% to 45.7%; the surviving M1 baseline records one pooled transport figure, 30.0% to 48.0%, which we compare with M2's **W** value. Under notation alone (**N**), computed for M2 only, retrieval rises from 14.4% to 30.7%. That 14.4% is the one pre-registered acceptance check that failed, the notation representation having been required to place source-side rank-1 retrieval in a 20–35% band; it is reported rather than tuned away, it belongs to a different representation from the 30.0% above and does not contradict it. Under **W**, **N** and **X**, on every pair and for both models, C_{auto} orders a verbatim copy below the model's rewrite and the rewrite below the held-out gold.

Sample-to-sample instability with the judge held fixed (M1, abstracting, vertical).

framing pair	$k=1$	$k=2$	$k=3$
vector spaces → module theory	14/15	8/15	11/15
probability → measure theory	15/15	12/15	10/15
set theory → category theory	10/15	0/15	1/15
pooled	**39/45**	**20/45**	**22/45**

Inter-rater reliability of the superseded first run. These figures come from the three same-family judges of the original run, not from the instrument used in §4; the primary code of that instrument transfers between judge families at $\kappa = 1.000$ and agrees with a blind expert reading at $\kappa = 0.900$ (§4.4). Krippendorff's α, computed on items with ≥ 2 raters: `equivalent` (PP binary) $\mathbf{0.807}$ ($n = 210$); `scope` (4-way) 0.599; e_2 0.538; `truth` 0.478; e_1 0.467; `flag` 0.436 (all $n = 914$ except truth, $n = 250$). Simplicity buys reliability, which is the argument for a judge-free compliance axis.

H Statistical procedure, pre-registration and decision rules

Statistical procedure. The k samples of an item are not exchangeable and rewrites within an item are not independent, so the unit of analysis throughout is the item. For each item we take the mean of its k codings in each direction and test the paired difference with an exact sign-flip permutation test, enumerating all 2^n sign assignments where $n \leq 20$ and sampling 5×10^4 otherwise. Confidence intervals are cluster bootstraps over items, stratified by framing pair where a pooled estimate is reported. Contingency comparisons use Fisher's exact test. Because one arm of the direction contrast is near zero, the paired test is arithmetically close to a one-sample test of the abstracting rate; we therefore report that rate with a binomial interval alongside every contrast rather than leaving the equivalence implicit.

I The regex–judge convergence on three populations

The same quantity, namely the rate at which a candidate that states the standing hypothesis is judged to have widened the domain against the rate for one that omits it, computed on each of the three populations in which it can be formed. **A**: Claude Opus 5 alone, both halves, 65 candidates, 7.1% against 90.2%. **B**: seven models, held-out items, 315 candidates, 2.1% against 85.9% (the value quoted in §4.4). **C**: three models on the ProofWiki items, 72 scorable candidates, where only 3 state the hypothesis at all and the comparison is uninformative.

J Additional experiments on Claude and GPT

This appendix reports a second set of experiments on two models, Claude Opus 5 and GPT-5.6 Sol, at $k = 3$ over all 90 items and all three axes. It covers analyses that the seven-model run does not: the lateral and functorial control axes, the per-pair typology, and the comparison of accounts of hypothesis presence. Its judged codes were produced under an earlier version of the scope codebook rather than the repaired one used in the main text, so its levels are not directly comparable with those of §4 and are read here for their internal contrasts.

Table 12: **Truth, content and scope under the bare prompt**: percentage of rewrites coded true, content-equivalent on the shared domain, and scope-preserving (`scope_ok`). M1 is coded by the adjudicated majority of three judges and M2 by one judge under the repaired codebook (§C.1); vertical cells pool $k = 3$ samples, and M2's control cells use one. n: M1 / M2.

			M1 (Claude Opus 5)			**M2 (GPT-5.6 Sol)**		
axis	**direction**	n	truth↑	equiv.↑	**scope↑**	truth↑	equiv.↑	**scope↑**
vertical	abstracting (promote)	135 / 135	98.5	99.3	**59.3**	98.5	98.5	**63.7**
vertical	specialising (absorb)	135 / 135	100	97.8	95.6	100	98.5	81.5
lateral	A→B	90 / 30	94.4	95.6	95.6	90.0	66.7	90.0
lateral	B→A	90 / 30	95.6	95.6	100	93.3	76.7	93.3
functorial	A→B	45 / 15	100	95.6	88.9	66.7	86.7	80.0
functorial	B→A	45 / 15	100	86.7	86.7	93.3	53.3	80.0

Table 13: **Abstracting scope preservation by framing pair** (%). *Canonical*: each model's registered coding, with a Fisher exact test. *Common judge*: both models' rewrites of byte-identical tasks re-coded blind to generator by one judge. *MELD gold*: reference translations ($n = 30$ per pair) under M2's judge.

	canonical coding			**common judge**		**reference**
framing pair, abstracting	M1↑	M2↑	p	M1↑	M2↑	MELD gold↑
vector spaces → modules	73.3	33.3	3×10^{-4}	33.3	33.3	96.7
probability → measure	80.0	64.4	0.16	66.7	66.7	43.3
set theory → categories	**24.4**	**93.3**	1.2×10^{-11}	**40.0**	**91.1**	83.3
pooled	59.3	63.7	0.53			74.4

RESULTS

These experiments proceed from phenomenon to mechanism to remedy. They locate the loss of scope (§J), characterise the failures by type and cost (§J) and, because both analyses rest on judged codes, use the judge-free measure to identify what governs hypothesis presence (§J) and test whether instruction restores directional control (§D.1). Uncertainties are one standard error and tests are exact and two-sided. The corresponding analyses in the present revision are reported in §4 under the repaired codebook, and the levels below are not comparable with them.

Content is preserved; generality is not If a model represents the content of a statement but not its level of generality, truth and content should survive translation, while scope should fail where a hypothesis must be introduced, namely when abstracting on vertical pairs, and Table 12 confirms this prediction. On vertical pairs, truth and content equivalence lie between 97.8 and 100% in every cell, whereas scope is preserved in $59.3\pm4.2\%$ of M1's abstracting rewrites and $63.7\pm4.1\%$ of M2's (Fisher exact test, $p = 0.53$), against 95.6% and 81.5% when specialising. The controls isolate the effect: with no generality to lose, scope is preserved in 80.0–100% of rewrites, so vertical abstraction is the only cell below 80%; the controls lose content instead (M2's equivalence falls to 53.3%), which an equivalence check detects. The vertical losses are silent: 40.7% of M1's and 36.3% of M2's abstracting rewrites change the domain without stating it, almost always by widening (102 of 107 and 55 of 60 non-identity domain codes are `sup`; Appendix G), and on the shared domain such a rewrite is identical to its source.

Since pooled rates may conceal heterogeneity and judge effects, Table 13 disaggregates them. Under canonical coding the endpoints reverse: M1 preserves scope best on vector spaces and worst on set theory (73.3 against 24.4%), and M2 shows the opposite pattern (33.3 against 93.3%). Re-coded by one judge, the models coincide on two pairs and differ only on set theory → categories (40.0 against 91.1%), so part of the canonical difference reflects the judge, which motivates the judge-free

Table 15: **Pin presence in both directions** (judge-free, $k = 3$): percentage of rewrites containing the registered pin. Promotion requires the pin (100% is correct), absorption its removal (0% is correct; a retained pin is redundant, not false); DSI in points. The item whose abstracting source contains the pin is excluded. Right: pin frequency in native prose over all 270 MELD pairs.

	M1 Claude Opus 5			M2 GPT-5.6 Sol			native prose	
framing pair	promote↑	absorb↓	DSI↑	promote↑	absorb↓	DSI↑	general	concrete
vector spaces ↔ modules	4.4%	82.2%	**−77.8**	17.8%	28.9%	−11.1	30/30	**0/30**
probability ↔ measure	20.0%	40.0%	−20.0	30.0%	26.7%	+3.3	10/30	**1/30**
set theory ↔ categories	33.3%	11.1%	+22.2	80.6%	13.9%	+66.7	25/30	**0/30**
pooled	**18.0%**	**47.7%**	**−29.7**	**41.4%**	**23.4%**	**+18.0**		

analysis of §J. The benchmark's reference translations preserve scope on 96.7% of vector-space pairs but only 43.3% of probability pairs, where gold-referenced scoring could not detect the relocation.

Failures relocate rather than falsify We next characterise the failures by type, which reflects the operation performed, and by cost, which determines whether outcome-based evaluation could detect them. On vertical pairs, failures are coded as a generalised conclusion (ΔC) and a coerced type (ΔT), with "vector space" read as "module", and rarely as an omitted hypothesis (ΔH: 3.1% on vector spaces ↔ modules; Table 5), as expected when the pin never appears in the source text. The remaining types are confined to the pairs whose structure admits them: referent drift occurs only on geometry ↔ commutative algebra (22 of 22 codes), the one anti-equivalence, where a dictionary applied covariantly yields the dual object (Appendix E), and illocutionary drift mainly on graph theory ↔ linear algebra (9 of 11).

With respect to cost (Table 14), both models preserve scope where omitting the pin would make the statement false (87.5% and 96.3%), and M2 where it would make it ill-formed (86.1%), but not where the widened statement remains true (40.0% and 33.3% under reference load; 58.0% and 40.9% under inert load). The loss is thus concentrated where no truth-checker, prover or equivalence score would register it, subject to three qualifications: the item-level crosswalk between the two annotations was not retained, ill-formed load is concentrated on set theory (eight of twelve items), and M2's reference cell rests on six rewrites. The cause of these failures is examined next with the judge-free measure.

Table 14: **Abstracting scope preservation by pin load.** Each model is scored under its own, independently produced `load` annotation (M1's has no ill-formed category); p: Fisher exact test, truth against the rest.

dropping the pin makes it…	M1↑	M2↑
false (*truth*)	**21/24** (87.5)	**26/27** (96.3)
ill-formed	n/a	31/36 (86.1)
true of other objects (*ref.*)	12/30 (40.0)	2/6 (33.3)
unchanged (*inert*, none)	47/81 (58.0)	27/66 (40.9)
p, truth vs. rest	0.0024	2.9×10^{-5}

Generality is inherited rather than derived A model that derived generality from the requirements of the translation would state the pin when promoting and omit it when absorbing (DSI $\approx +100$); one that copied the source's surface would yield a negative DSI. Table 15 measures both directions without a judge. **M1's DSI is negative**: pooled, it states the pin in $18.0 \pm 3.6\%$ of rewrites where it is required and $47.7 \pm 4.7\%$ where it is redundant. On vector spaces and modules, it states the field hypothesis in 82.2% of rewrites *into vector-space prose*, where the noun encodes it, and 4.4% *into module-theory prose*, where it is required, a nineteen-fold difference in the wrong direction; paired within item, 5 items carry the pin only when abstracting against 16 only when specialising (exact McNemar test, $p = 0.027$). M2's pooled DSI is positive (+18.0; $41.4 \pm 4.7\%$ against $23.4 \pm 4.0\%$) but again describes no single pair: it controls the pin on set theory ↔ categories (+66.7) alone.

Three accounts of pin presence. The divergent per-pair profiles suggest different policies. We fit three accounts by binomial likelihood with one lapse rate (Table 16; Appendix G): *requirement*, under which the pin is always stated when promoting and never when absorbing; *source inheritance*,

Table 16: **Three accounts of pin presence** (Table 15 counts): log-likelihood relative to each model's best account and fitted lapse ε (0.50: uninformative).

	M1		M2	
account	$\Delta \log L \uparrow$	$\varepsilon \downarrow$	$\Delta \log L \uparrow$	$\varepsilon \downarrow$
requirement	-21.2	0.50	-3.0	0.41
source inheritance	**0**	0.25	-6.7	0.50
target register	-21.2	0.50	**0**	0.36

Table 17: **Judge-free mechanical axes.** AUC: leave-one-out separation of a pair's two framings (range over pairs). Twin: retrieval of, and cosine to, the cross-framing twin (Appendix G).

	M1	M2
AUC, words (**W**)	.894–1	.951–1
AUC, notation (**N**)	**.909–1**	**.921–.998**
AUC, char. n-grams (**X**)	.902–1	.962–1
twin rank-1 (%)↑	30.0 → 48.0	26.1 → 45.7
twin cosine↑	.149 → .348	.145 → .333

under which the source's pin state is copied; and *target-register imitation*, under which the pin is stated as often as native prose in the target dialect states it. Source inheritance describes M1 decisively, leading both alternatives by 21.2 in log-likelihood, while both fit only at chance ($\varepsilon = 0.50$). Target-register imitation best describes M2, ahead of requirement by 3.0 and of inheritance by 6.7, a modest, descriptive margin over three pairs; consistently, its one controlled pair is the one whose pin, "In $\mathbf{Set}$", is native category-theoretic vocabulary, whereas module theory opens with "let R be a ring", which the pin must override. For neither model is the requirement the best account.

Surface recoverability of framings. Target-register imitation presupposes that a framing's register is recoverable from its surface form, as Table 17 confirms: with every multi-letter word removed, the two framings of a pair separate at leave-one-out AUC of 0.909–1.000 (M1) and 0.921–0.998 (M2), above a pre-registered 0.90. Translation moves a statement toward its cross-framing twin but leaves it more than half short in both models (Appendix G).