EXECUTABLE TRUST: QUERY-BLIND VALIDATION OF RULE PROPOSALS FOR LINGUISTICS OLYMPIAD REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Linguistics Olympiad problems require solvers to induce latent rules from sparse evidence in unfamiliar languages. Yet final-answer accuracy is fundamentally insufficient: fluent but invalid rules routinely reach the same prediction, making it impossible to distinguish supported analyses from plausible confabulations. We introduce Executable Trust, a generator-agnostic framework that compiles rule proposals from language models or human experts into typed executable hypotheses and validates them before query labels are revealed. Four gates—leave-one-context reconstruction, counterexample sensitivity, perturbation stability, and support calibration—jointly produce an auditable admit, reject, or abstain decision with a replayable evidence trace. Across six rule families spanning numeral systems, morphology, allomorphy, agreement, role alignment, and conditioned forms, the protocol decisively admits three supported families (Yidiny morphology, Sulka numeral-classifier structure, and Zoque conditioned allomorphy) and rejects three under-specified or unstable alternatives (Kayapo, Wambaya, and Benabena), with every decision backed by reconstruction scores, perturbation variance, and counterexample counts. On 32 query-blind typed queries, a strict selective policy answers 20 and is correct on all 20—yielding 62.5% coverage at 100% answered precision (95% Wilson interval: 83.9–100.0%). The 12 abstentions are not failures but calibrated refusals, each traced to reconstruction or stability gates that flagged insufficient evidence. An expert-authored derivation language of 21 rules executes all hidden-context analyses correctly under systematic source-side ciphers across 5 cipher strengths and rejects 68 of 71 corrupted controls (95.8% rejection rate); the three survivors are behaviorally equivalent rule-order permutations. Executable validation constitutes a falsifiable trust layer between linguistic hypothesis generation and answer production, forcing abstention and representation failures into the open rather than allowing them to hide behind answer accuracy.

1 INTRODUCTION

Linguistics Olympiad (LO) puzzles ask a solver to recover latent morphological, syntactic, semantic, phonological, or numeral structure from a small set of examples in an unfamiliar language. They are therefore an appealing test of sparse induction rather than ordinary language familiarity. Recent benchmarks confirm both their difficulty and their sensitivity to presentation: LingOly, Linguini, IOLBENCH, LOBSTER, and LingBench++ cover low-resource, extinct, and constructed settings Bean et al. (2024); Sánchez et al. (2024); Goyal & Dan (2025); Lian et al. (2025b;a), while LingOly-TOO, ModeLing, and Rosetta-to-Match-Up show that obfuscation, templatisation, and puzzle format can change measured performance Khouja et al. (2025); Chi et al. (2024); Majmudar et al. (2026).

Final-answer accuracy, however, conflates two events: proposing a valid analysis and reaching a correct output. A fluent explanation can fit one row, silently contradict another, and still guess the query. Conversely, a partially supported rule may be useful only for a restricted query regime. Neither a language model nor a human expert should be trusted solely because the proposal is articulate. What is missing is an intermediate object that can fail before it is allowed to answer.

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.

We introduce *Executable Trust*, a generator-agnostic interface that compiles a proposed linguistic rule into a typed, executable hypothesis. The validator does not inspect query labels while deciding whether the hypothesis has earned admission. It instead executes four gates: leave-one-context reconstruction, counterexample sensitivity, perturbation stability, and support calibration. The output is an admit, reject, or abstain decision plus a replayable evidence trace. Proposal generation can improve or change model families without changing this trust contract.

The registered study begins with a six-family pilot. Three expert-prior hypotheses are admitted, three weak alternatives are rejected, and a strict policy answers 20 of 32 typed queries correctly while abstaining on the remainder. We then freeze the validator and perform a substantially larger confirmation: 57,600 proposals from eight generators, 2,400 cluster-disjoint queries, twelve rule families, 2,500 expert derivations, 1,000 frozen adversarial confirmations, and a 1,200-trace three-linguist audit. At the frozen operating point, the 32B tier attains 56.21% coverage and 93.33% answered precision; the 70B tier reaches 61.83% and 93.87%, respectively. The confirmation escape rate is 2.60%, and expert accepted-rule precision is 96.54%. These results do not turn the validator into an unrestricted LO solver. They test the narrower, reusable claim that unsupported analyses can be blocked and supported ones can be used selectively.

Our contributions are:

- a proposal-source-independent contract for converting linguistic analyses into falsifiable executable hypotheses;
- four query-blind gates that distinguish rejection from calibrated abstention and preserve a decision-level evidence trace;
- a pilot instantiation over six typed families and an expert-authored derivation DSL with positive and corrupted controls; and
- a frozen, cluster-disjoint confirmation over 57,600 model proposals that measures scale, risk–coverage, adaptive attacks, expert controls, and human agreement under one validator.

2 Related Work

Olympiad reasoning. LO benchmark construction has accelerated through LingOly, Linguini, IOLBENCH, LOBSTER, and LingBench++, alongside solve-and-generate and teaching-oriented work Bean et al. (2024); Sánchez et al. (2024); Goyal & Dan (2025); Lian et al. (2025b;a); Majmudar & Filatova (2025); Sakunkoo & Sakunkoo (2025). Error analyses emphasize compositional inference, sparse evidence, and the fragility of surface-form reasoning Choudhary et al. (2025); Lin et al. (2025). We evaluate a different object from a benchmark score: whether a proposed analysis itself survives tests that do not use the target query label.

Search and verification. Chain-of-thought, self-consistency, least-to-most prompting, Tree of Thoughts, inductive linguistic prompting, and IOL-specific inference-time scaling increase candidate supply or search budget Wei et al. (2022); Wang et al. (2022); Zhou et al. (2022); Yao et al. (2023); Ramji & Ramji (2024); Garnham & Shareghi (2026). Program-of-thought, program search, neurosymbolic inference, program verifiers, and verifier-guided self-correction instead demonstrate the value of executable checks Chen et al. (2022); Romera-Paredes et al. (2024); Olausson et al. (2023); Toh et al. (2024); Zhang et al. (2024). Executable Trust connects these ideas to typed linguistic operations such as allomorphy, agreement slots, role alignment, numeral classifiers, and correspondence tables. Its distinctive unit is a selective trust decision: a proposal may be rejected globally, admitted but withheld for an unsupported query, or admitted and used.

3 Executable Trust

3.1 Typed Proposal Contract

Let a puzzle provide visible context $C = \{(x_i, y_i)\}_{i=1}^{m}$ and a query q. A proposal source g emits a candidate analysis z. A family compiler maps z to an executable hypothesis h with typed objects, permitted operations, required evidence, and a deterministic executor:

$$A(h, C, q) = \mathrm{execute}(h, C, q).$$

Unlike a free-form rationale, h must reconstruct observations and expose which rules and rows support each candidate. A trace records the source and model family, hypothesis family and variant, held-out context row, parser status, reconstruction outcome, counterexamples, perturbation outcomes, support score, and final decision. Expert-authored and model-generated proposals share the validator but retain distinct provenance.

The current library contains twelve families: numeral classifiers, morphology, allomorphy, agreement slots, correspondence tables, role alignment, conditioned forms, phonological rules, suffix chains, inventory assignment, case marking, and reduplication. These are bounded hypothesis spaces, not a universal grammar or automatic inventory discovery claim.

3.2 Four Query-Blind Gates

Reconstruction. For each context row r, the validator removes r, executes the proposal using C_{-r}, and checks whether the hidden row is recovered:

$$C_{-r} = C \setminus \{r\}, \qquad \widehat{r} = \mathrm{select}(A(h, C_{-r}, x_r), C_{-r}).$$

The pilot admits a family when either selected or fixed top-5 reconstruction is at least 0.50. Malformed proposals and reconstruction failures remain in the denominator.

Counterexamples and perturbations. A proposal must reject rule-violating controls and remain decision-stable under applicable row permutations, distractor insertions, and systematic source-side ciphers. Perturbations preserve the linguistic target while changing irrelevant representation. Failure rejects the variant rather than being silently repaired after query inspection.

Support calibration. An admitted hypothesis need not support every query. The pilot answers only when selected reconstruction is at least 0.75 over at least eight applicable calibration rows. The confirmation generalizes this rule to per-query support. For query q, let s_{qr} be support from relevant runs. With the threshold frozen at $\tau = 0.75$,

$$\mathrm{answer}(q) = \mathbb{1}\left[\frac{1}{|R_q|}\sum_{r \in R_q} s_{qr} \geq \tau\right].$$

No oracle router is used. The correctness attached to an answered query is the latent correctness of the highest-support run, with a deterministic flat-index tie break; latent correctness is defined for every query–run before the answer/abstain decision. Thus changing τ changes coverage, not the underlying correctness labels.

Decision semantics. *Reject* means that a proposal or variant fails an executable gate. *Abstain* means an admitted hypothesis lacks enough support for this query. *Admit* enables selective answering but is not itself a claim of universal correctness. This separation is central: it prevents unsupported coverage from being credited as evidence of trust.

Compile, reconstruction, and stress-test failures are properties of a proposal or variant. Calibration failures are query-specific and therefore produce abstention rather than retroactively rejecting an otherwise supported family. The final audit never changes a decision; it only assigns its outcome.

3.3 Executable Derivations

The Teacher Derivation DSL turns an expert analysis into a sequence of typed operations with explicit inputs, intermediate states, and failure conditions. The 21-row pilot covers eight languages, ten operation types, and 35 declared failure modes. Each derivation is executed on a hidden context row, under five systematic cipher strengths, and against corrupted controls. The DSL is expert-authored supervision; it evaluates executability and does not claim to discover a complete derivation automatically.

Family	Selected	Top-5	Decision
Yidiny morphology	44/92	60/92	admit
Sulka numerals	48/68	48/68	admit
Zoque allomorphy	54/80	54/80	admit
Kayapo prototype	4/32	4/32	reject
Wambaya prototype	2/68	4/68	reject
Benabena prototype	0/60	8/60	reject

Table 1: Pilot leave-one-context reconstruction. The pre-query-label gate admits when selected or fixed top-5 reconstruction is at least 0.50.

Variant	Baseline	Full cipher
Yidiny hardcoded	44/60/92	0/0/92
Sulka induced marker	48/48/68	48/48/68
Zoque suffix chain	54/54/80	14/20/80

Table 2: Pilot robustness by selected/top-5/total reconstruction. The cipher test promotes Sulka and blocks a broad robustness claim for the other two admitted families.

4 Evaluation

4.1 Registered Pilot

The pilot uses public Linguini contexts Sánchez et al. (2024). Six family prototypes cover numerals, morphology, allomorphy, agreement, role alignment, and conditioned forms. Family definitions, the 0.50 admission rule, variant tests, the 0.75 support rule, and the minimum support count are fixed before query labels are inspected. The 32 typed-query evaluation is a query-blind development diagnostic, not a held-out benchmark score. The separate 21-row derivation study uses expert-authored annotations over public contexts rather than student data.

4.2 Frozen Confirmation

Proposal matrix. We sample 300 cluster-disjoint public Olympiad problems with eight queries per problem. Eight instruction-tuned generators span four size tiers and four model families: Qwen2.5-7B, Llama-3.1-8B, Qwen2.5-14B, Mistral-NeMo, Qwen2.5-32B, Yi-1.5-34B, Qwen2.5-72B, and Llama-3.1-70B. Each generator emits eight independently decoded proposals per problem under seeds 17, 42, and 83, for $8 \times 300 \times 8 \times 3 = 57{,}600$ traces. Repository identifiers and the complete scale and generator tables are supplied in the supplement. Models are proposal sources; no model bypasses the deterministic validator.

Transfer and controls. The held-out evaluation contains 2,400 unique queries from the 300 clusters, excluded from proposal development. The expanded library has twelve rule families. A separate set of 2,500 expert derivations measures parsing, admission, answered precision, cipher robustness, and rejection of corrupted controls. Adversarial development and confirmation each contain 1,000 records. The development set may reveal failure modes; all recovery rules are frozen before opening the confirmation set. A stratified human audit samples 1,200 accepted, rejected, and escaped traces. Three unpaid volunteers with prior Linguistics Olympiad participation judge executability, evidential support, answer preservation, and the validator decision independently before majority aggregation. Each volunteer rates all 1,200 traces under a common rubric while blind to generator

Policy	Top-1	Answered	Correct
Default support	26/32	21/32	20/21
Visible-fit guard	26/32	20/32	20/20

Table 3: Pilot selective answering. Top-1 is the ungated diagnostic; answered correctness is evaluated only after the support policy is fixed.

Component	Design	Units
Six-family pilot	6 families	32 queries
Teacher DSL pilot	21 derivations	71 corruptions
Model confirmation	300 clusters	57,600 proposals
Held-out transfer	300 clusters	2,400 queries
Expert confirmation	12 families	2,500 derivations
Adversarial dev.	–	1,000 attacks
Adversarial confirm.	–	1,000 attacks
Blind human audit	–	1,200 traces

Table 4: Study components. Development attacks diagnose failures; only the post-freeze confirmation set supports the reported attack claim.

Analysis	Experimental unit	Dependence handling	Primary role
Proposal matrix	generator–problem–decode–seed	complete 57,600-trace denominator	supply and gate flow
Held-out transfer	query within problem cluster	cluster/query hierarchical bootstrap	coverage and precision
Adversarial confirmation	independently generated attack record	Wilson interval over 1,000 records	escape rate
Expert confirmation	expert derivation within family	family-wise results plus aggregate	source-independent validity
Human audit	trace with three independent ratings	agreement plus Fleiss' κ	decision interpretability

Table 5: Experimental units and inferential roles. Proposal traces are not miscounted as 57,600 independent query outcomes.

identity and the held-out query label. They were informed that de-identified task labels would be used for research and participated voluntarily; no names, demographics, or sensitive data were retained. No formal institutional ethics review or exemption determination was obtained because the task evaluated proposal traces rather than personal participant outcomes.

Inference. Coverage and answered precision are computed per unique query, not per proposal. We obtain 95% intervals from 5,000 hierarchical bootstrap replicates that resample 300 problem clusters and then eight queries within a cluster. Wilson intervals are used only for the 1,000 independent attack records. We report parse rate, reconstruction rate, admitted count, false admission among admitted proposals, coverage, precision, expected calibration error (ECE), and Brier score. Support is a routing score rather than a declared probability, so answered-region ECE is the more relevant diagnostic.

This separation prevents a common form of pseudo-replication. Eight generators and three seeds create many proposals for the same problem, but the reported transfer estimand remains a decision over 2,400 unique queries. Likewise, the 300 problem clusters, rather than individual proposal traces, anchor the bootstrap. The attack set is generated as independent records and therefore uses a binomial interval. Expert and human analyses answer different questions: the former checks a controlled high-quality proposal source, while the latter tests whether decision traces match linguists' reading of the protocol.

Freeze chronology and leakage control. Table 4 separates the registered pilot, development diagnostics, and confirmation evidence. The pilot fixes decision semantics and establishes that the interface can express both positive and negative outcomes. The adversarial development set may be inspected to identify parser and gate failures. Recovery logic is then frozen, and the 1,000 confirmation attacks are opened once. The 2,400 transfer-query labels are not inputs to proposal parsing, reconstruction, perturbation, or support aggregation. Threshold sweeps are reported after the frozen $\tau = 0.75$ result and do not replace it.

Estimands. For a size tier t, proposal validity and reconstruction are computed over all 14,400 proposal traces. False admission is

$$\mathrm{FA}_t = \frac{\#\{\text{admitted and incorrect proposals}\}}{\#\{\text{admitted proposals}\}}.$$

For the unique query set Q, the selective estimands are

$$\mathrm{Cov}_t = \frac{\sum_{q \in Q} a_{tq}}{|Q|}, \qquad \mathrm{Prec}_t = \frac{\sum_{q \in Q} a_{tq} c_{tq}}{\sum_{q \in Q} a_{tq}},$$

where a_{tq} is the frozen answer indicator and c_{tq} is latent correctness. Reporting both terms prevents a high-precision, vanishing- coverage system from looking uniformly reliable. Attack escape is a separate estimand because an attack is an attempt to produce an incorrect admitted decision, not an ordinary held-out query.

5 Results

5.1 Pilot Establishes the Contract

The same answer-free gate admits Yidiny, Sulka, and Zoque while rejecting the three weak prototypes (Table 1). Yidiny illustrates why the fixed top-5 arm is specified in advance: 60/92 candidates reconstruct the hidden row even though selected reconstruction is 44/92. On the 32 typed queries, the ungated top-1 score is 26/32. The default support policy answers 21 with 20 correct; the stricter visible-fit guard answers 20 and is correct on all 20 (62.5% coverage; 95% Wilson precision interval 83.9–100.0%). The twelve withheld items are observable abstentions, not dropped errors (Table 3).

Admission alone is insufficient for a robustness claim. Sulka's induced marker variant exactly preserves its 48/68 reconstruction under the full source cipher. Yidiny's hardcoded representation collapses under the same test, and Zoque recovers only 14 selected and 20 top-5 rows out of 80 (Table 2). The protocol therefore promotes the Sulka variant but retains Yidiny and Zoque as representation-boundary cases. This negative result is operational: without the perturbation gate, both variants would be credited solely from their clean reconstruction.

The 21 expert-authored derivations all execute on their hidden context rows and pass strict support. Every row also executes under each of five source-side cipher strengths. Corruption controls are rejected in 68/71 cases; the three survivors are behaviorally equivalent rule-order permutations. The corresponding query-generation scope test is deliberately negative (0/138 replay and 1/138 library reconstruction), separating derivation verification from automatic rule discovery.

5.2 Negative Evidence Distinguishes Supply from Trust

The fixed pilot also tests whether a fluent rationale or a larger candidate pool is sufficient. It is not. Direct, chain-of-thought, compact-hypothesis, and shallow-verifier routes solve none of the 50 local items, whereas the executable hybrid solves 17. For Benabena, a wide top-20 oracle contains 20/60 hidden answers, but the predeclared selector recovers none and the best coarse cross-sibling policy recovers only 8/60. The family is therefore rejected. Similarly, improving Yidiny sibling reconstruction from 7/23 to 13/23 does not alter real selective-query predictions. Crediting oracle membership would erase exactly the selection failure that the trust layer is designed to expose. These are diagnostics of covered components and candidate supply, not held-out confirmation estimands.

5.3 Scale Expands Coverage without Bypassing Validation

Parse and reconstruction improve monotonically with scale (Table 6), yielding more admitted proposals and broader selective coverage. Yet false admission remains between 2.38% and 3.06%, rather than growing with the admitted pool. At the frozen operating point, answered precision increases from 86.33% at 7B to 93.87% at 70B while coverage rises from 36.88% to 61.83%. The Qwen-only curve shows the same pattern within one model family: coverage is 45.75%, 54.38%, 61.25%, and 65.50% from 7B through 72B, with precision 87.07%, 90.50%, 92.72%, and 93.58%. The full per-generator table shows that the aggregate is not solely a cross-family artifact (Table 7). Within every

Tier	Parse	Recon.	Admitted	False adm.	Coverage (95% CI)	Precision (95% CI)
7B	42.01%	17.98%	1,832	3.06%	36.88% [34.25, 39.50]	86.33% [83.12, 89.37]
14B	52.72%	25.92%	2,774	2.38%	43.92% [41.17, 46.75]	90.80% [88.44, 93.10]
32B	62.71%	33.47%	3,518	2.76%	56.21% [53.54, 58.88]	93.33% [91.45, 95.06]
70B	71.11%	40.32%	4,282	2.55%	61.83% [59.08, 64.50]	93.87% [92.12, 95.53]

Table 6: Frozen confirmation by size tier. Parse, reconstruction, admission, and false-admission use 14,400 proposals per tier. Coverage and precision use 2,400 cluster-disjoint queries with hierarchical-bootstrap intervals.

Generator	Tier	Parse	Coverage	Answered precision
Qwen2.5-7B	7B	43.15%	1,098/2,400 (45.75%)	87.07%
Llama-3.1-8B	7B	40.88%	907/2,400 (37.79%)	78.61%
Qwen2.5-14B	14B	55.18%	1,305/2,400 (54.38%)	90.50%
Mistral-NeMo	14B	50.26%	1,054/2,400 (43.92%)	85.48%
Qwen2.5-32B	32B	65.56%	1,470/2,400 (61.25%)	92.72%
Yi-1.5-34B	32B	59.86%	1,294/2,400 (53.92%)	89.88%
Qwen2.5-72B	70B	72.92%	1,572/2,400 (65.50%)	93.58%
Llama-3.1-70B	70B	69.31%	1,422/2,400 (59.25%)	91.21%

Table 7: Generator-level confirmation at the same frozen support threshold. Coverage and precision are computed over the shared 2,400 queries.

tier, the second generator has lower coverage and precision than Qwen, yet the validator retains substantial selective precision: 78.61% for Llama-3.1-8B, 85.48% for Mistral-NeMo, 89.88% for Yi-1.5-34B, and 91.21% for Llama-3.1-70B. This heterogeneity tests the contract under materially different proposal supplies rather than averaging replicas of one model line.

These trends support the interface claim rather than a model leaderboard. Larger generators provide more parseable and reconstructing proposals; the same validator continues to reject malformed and unsupported ones. The proposal denominator includes every generator, seed, and decoding, including failed parses and abstentions.

5.4 Support Produces an Auditable Risk–Coverage Tradeoff

Table 8 exposes the cost of abstention instead of hiding it. Raising τ from 0.75 to 0.82 increases precision by 2.16 points but reduces coverage by 23.88 points. Lowering it to 0.68 gains 17.79 coverage points at a 1.94-point precision cost. The frozen point therefore represents a declared operating choice, not the best row selected after seeing labels. Full-set ECE is 0.1065, answered-only ECE is 0.0972, and the latent Brier score is 0.1226. A label-permutation sanity check gives $p < 10^{-4}$: observed precision is 93.33%, whereas the permuted mean is $85.68\% \pm 0.65\%$.

τ	Answered	Coverage	Precision
0.92	97	4.04%	98.97%
0.87	382	15.92%	97.91%
0.82	776	32.33%	95.49%
0.75	**1,349**	**56.21%**	**93.33%**
0.68	1,776	74.00%	91.39%
0.60	2,107	87.79%	88.70%
0.52	2,294	95.58%	87.05%
0.44	2,361	98.38%	86.11%

Table 8: Risk–coverage sweep for the 32B tier. The bold operating point was frozen before confirmation; correctness labels are unchanged across rows.

Attack type	n	Escape	Rate (95% CI)
Gate-rule violation	217	1	0.46 [0.08, 2.56]
Parser boundary	183	2	1.09 [0.30, 3.90]
Cipher corruption	204	5	2.45 [1.05, 5.61]
Support inversion	198	4	2.02 [0.79, 5.08]
Adaptive generation 1	87	6	6.90 [3.20, 14.24]
Adaptive generation 2	64	4	6.25 [2.46, 15.00]
Adaptive generation 3	47	4	8.51 [3.36, 19.93]
Total	**1,000**	**26**	**2.60 [1.78, 3.78]**

Table 9: Frozen adversarial confirmation. Intervals are Wilson intervals over independent attack records.

Rule family	n	Esc.	Rate	Rule family	n	Esc.	Rate
Numeral classifiers	84	2	2.38%	Conditioned forms	72	1	1.39%
Morphology	103	4	3.88%	Phonological rules	81	2	2.47%
Allomorphy	78	1	1.28%	Suffix chains	76	6	7.89%
Agreement slots	94	5	5.32%	Inventory assignment	64	2	3.12%
Correspondence tables	87	1	1.15%	Case marking	92	2	2.17%
Role alignment	83	0	0.00%	Reduplication	86	0	0.00%

Table 10: Adversarial confirmation by linguistic rule family. The supplement reports Wilson intervals for every row; suffix chains are the worst family.

5.5 Frozen Attacks Reveal Both Strength and Boundary

Only 26 of 1,000 confirmation attacks escape the validator (Table 9). Conventional gate, parser, cipher, and support violations remain below 2.5%; adaptive generation is harder, reaching 8.51% in generation 3, although its $n = 47$ interval is wide. By linguistic family, suffix chains are the worst case at 6/76 (7.89%, 95% CI 3.67–16.17); role alignment and reduplication have no observed escapes, with upper confidence bounds of 4.42% and 4.28%. Reporting the worst family prevents the aggregate 2.60% from concealing a concentrated weakness (Table 10). Agreement slots are the next-largest family rate at 5.32%, while the remaining nonzero families range from 1.15% to 3.88%. The pattern identifies suffix composition and adaptive proposal generation as priorities for future gate design rather than implying uniform robustness.

The development set exposed a role-alignment failure and motivated a frozen recovery rule. Its role-family escape rate changes from 2/83 (2.41%) on the post-recovery development set to 0/83 on confirmation. The difference is 2.41 points with a 95% interval of $[-0.89, 5.71]$, so we treat it as descriptive evidence, not a causal improvement claim.

Rule family	n	Parse	Accept	Precision (correct/accepted)	Corrupt reject
Numeral classifiers	214	98.13%	93.46%	98.50% (197/200)	97.20%
Morphology	221	94.57%	89.14%	96.95% (191/197)	98.19%
Allomorphy	198	94.95%	85.86%	98.24% (167/170)	95.96%
Agreement slots	217	95.39%	88.48%	96.35% (185/192)	96.31%
Correspondence tables	203	98.52%	93.60%	96.84% (184/190)	96.06%
Role alignment	196	94.90%	85.71%	93.45% (157/168)	94.90%
Conditioned forms	209	97.13%	91.87%	96.35% (185/192)	97.61%
Phonological rules	211	95.26%	89.57%	97.35% (184/189)	97.16%
Suffix chains	207	97.58%	91.79%	96.84% (184/190)	97.58%
Inventory assignment	201	97.51%	95.52%	95.83% (184/192)	97.51%
Case marking	213	97.18%	91.08%	98.45% (191/194)	95.77%
Reduplication	210	93.33%	84.76%	92.70% (165/178)	95.24%
All	**2,500**	**96.20%**	**90.08%**	**96.54% (2,174/2,252)**	**96.64%**

Table 11: Expert-proposal confirmation by rule family. Every family is kept in the denominator, including parse failures and rejected derivations.

Audit outcome	Count	Rate
Expert derivations parsed	2,405/2,500	96.20%
Expert derivations admitted	2,252/2,500	90.08%
Correct among admitted	2,174/2,252	96.54%
Corrupted controls rejected	2,416/2,500	96.64%
Cipher drop, family-weighted	–	2.19 pp
Human decision agreement	1,107/1,200	92.25%
Human inter-annotator agreement	–	$\kappa = 0.837$

Table 12: Independent expert and human-facing confirmation. Human agreement compares the validator decision with the majority-aggregated judgment.

Cipher level	Recon. pass	Drop from plaintext
Plaintext	486/500 (97.20%)	0.00 pp
Caesar shift	481/500 (96.20%)	1.00 pp
Substitution	480/500 (96.00%)	1.20 pp
Transposition	475/500 (95.00%)	2.20 pp
Compound	478/500 (95.60%)	1.60 pp

Table 13: Expert reconstruction under five equally sized cipher levels.

5.6 Expert and Human Audits Support Decision Validity

Expert controls provide a proposal source that is independent of model scale (Tables 12 and 11). Among 2,252 admitted expert derivations, 2,174 are correct (96.54%). Corrupted-control rejection is 96.64%. Every one of the twelve families exceeds 92.7% answered precision; reduplication is the lowest at 92.70% and numeral classifiers the highest at 98.50%. Across five equally sized cipher levels, reconstruction changes from 97.20% in plaintext to 96.20%, 96.00%, 95.00%, and 95.60%, a mean drop of 1.50 points (Table 13). The separate family-weighted drop is 2.19 points. Thus robustness is not carried only by one easy cipher or one family.

The majority-aggregated blind human judgment agrees with the validator outcome on 1,107/1,200 traces and attains Fleiss' $\kappa = 0.837$. This does not make human judgment an oracle for every linguistic analysis. It shows that the machine-readable admit/reject/escape decisions usually correspond to the criteria linguists were asked to apply, while leaving 93 disagreements for error analysis.

6 Discussion

7 Conclusion

Executable Trust inserts a falsifiable layer between linguistic rule proposal and answer production. The registered six-family pilot shows how reconstruction, perturbation, and support gates yield admit, reject, and abstain decisions. The frozen confirmation extends the same contract to 57,600 proposals, 2,400 cluster-disjoint queries, twelve families, expert controls, adaptive attacks, and human review. Larger generators increase coverage without bypassing the validator, while explicit abstention preserves an inspectable risk–coverage tradeoff. The result is not a universal LO solver but a reusable answer to a narrower question: when has a generated or expert linguistic analysis earned the right to answer?

AI Use Statement

AI assistance was used for code and manuscript-development support. The authors are responsible for all claims, references, numerical results, and artifact contents; no AI system is an author or cited as an evidence source.

REFERENCES

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages, 2024. URL https://arxiv.org/abs/2406.06196.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2022.

Nathan A. Chi, Teodor Malchev, Riley Kong, Ryan A. Chi, Lucas Huang, Ethan A. Chi, R. Thomas McCoy, and Dragomir Radev. Modeling: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of SIGTYP*, 2024. URL https://arxiv.org/abs/2406.17038.

Mukund Choudhary, KV Aditya Srivatsa, Gaurja Aeron, Antara Raaghavi Bhattacharya, Dang Khoa Dang Dinh, Ikhlasul Akmal Hanif, Daria Kotova, Ekaterina Kochmar, and Monojit Choudhury. Unveiling: What makes linguistics olympiad puzzles tricky for llms?, 2025. URL https://arxiv.org/abs/2508.11260.

Jamie Garnham and Ehsan Shareghi. Could language models win the international linguistics olympiad? In *Proceedings of the 2026 Conference on Computational Natural Language Learning*, 2026. URL https://openreview.net/forum?id=dioc2KgOKg.

Satyam Goyal and Soham Dan. Iolbench: Benchmarking llms on linguistic reasoning, 2025. URL https://arxiv.org/abs/2501.04249.

Jude Khouja, Karolina Korgul, Simi Hellsten, Lingyi Yang, Vlad Neacşu, Harry Mayne, Ryan Kearns, Andrew Bean, and Adam Mahdi. Lingoly-too: Disentangling memorisation from reasoning with linguistic templatisation and orthographic obfuscation, 2025. URL https://arxiv.org/abs/2503.02972.

Da-Chen Lian, Ri-Sheng Huang, Pin-Er Chen, Chunki Lim, You-Kuan Lin, Guan-Yu Tseng, Zhen-Yu Lin, Pin-Cheng Chen, and Shu-Kai Hsieh. Lingbench++: A linguistically-informed benchmark and reasoning framework for multi-step and cross-cultural inference with llms, 2025a. URL https://arxiv.org/abs/2507.16809.

Da-Chen Lian, Ri-Sheng Huang, Pin-Er Chen, Chunki Lim, You-Kuan Lin, Guan-Yu Tseng, Zhen-Yu Lin, Pin-Cheng Chen, and Shu-Kai Hsieh. LOBSTER: Linguistics olympiad benchmark for structured evaluation on reasoning. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pp. 193–229, National Taiwan University, Taipei City, Taiwan, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-379-1. URL https://aclanthology.org/2025.rocling-main.23/.

Zheng-Lin Lin, Yu-Fei Shih, and Shu-Kai Hsieh. Probing large language models in reasoning and translating complex linguistic puzzles, 2025. URL https://arxiv.org/abs/2502.00817.

Neh Majmudar and Elena Filatova. Can llms solve and generate linguistic olympiad puzzles?, 2025. URL https://arxiv.org/abs/2509.21820.

Neh Majmudar, Anne Huang, Jinfan Frank Hu, and Elena Filatova. From rosetta to match-up: A paired corpus of linguistic puzzles with human and llm benchmarks, 2026. URL https://arxiv.org/abs/2605.13408.

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger P. Levy. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Raghav Ramji and Keshav Ramji. Inductive linguistic reasoning with large language models, 2024. URL https://arxiv.org/abs/2412.17819.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL https://doi.org/10.1038/s41586-023-06924-6.

Jonathan Sakunkoo and Annabella Sakunkoo. Lingdex.org:leveraging LLMs to structure and explore linguistic olympiad puzzles for learning and teaching linguistics. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pp. 1053–1057, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2025.ranlp-1.121/.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. Linguini: A benchmark for language-agnostic linguistic reasoning, 2024. URL https://arxiv.org/abs/2409.12126.

Vernon Y. H. Toh, Deepanway Ghosal, and Soujanya Poria. Not all votes count! programs as verifiers improve self-consistency of language models for math reasoning, 2024.

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, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2022.

A Trace-Level Diagnosis and Interface Use

What the confirmation changes. The pilot establishes that a query-blind executable contract can admit and reject concrete families and can use abstention to protect answered precision. The confirmation tests whether that behavior survives a larger proposal supply, more rule families, cluster-disjoint queries, adaptive attacks, and independent expert and human checks. It does: coverage expands with proposal scale, false admission remains near 2–3%, and the frozen threshold yields over 93% answered precision at both 32B and 70B. This is stronger than selecting a single successful pilot, because every failed parse, rejection, and abstention remains visible in the denominator.

Limitations. The executable library and derivation DSL are expert-designed, so the method validates proposals rather than discovering an unrestricted linguistic theory. The public puzzle sources make contamination possible for proposal generators; cluster separation protects the evaluation protocol but does not turn the results into a claim of unseen-language model knowledge. The eight generators were chosen to span sizes and families, not to rank current frontier systems. Support is a routing score rather than a calibrated probability, and the ECE values reflect that distinction. Adaptive generation and suffix chains remain the clearest empirical weaknesses.

Why query blindness matters. Gate decisions use context reconstruction, controlled corruption, and support computed without query labels. Labels are opened only to evaluate the fixed decision. The risk–coverage sweep is diagnostic around the declared $\tau = 0.75$ point, not a retrospective threshold search. This ordering makes the central claim falsifiable: a proposal can fail before a correct final answer is observed.

Responsible release boundary. The anonymous artifact reports repository identifiers, seeds, frozen decision rules, aggregate and family-wise tables, and a verifier for paper-facing arithmetic. Public puzzle rows retain source and license fields. Human-facing release material is limited to de-identified decisions and agreement statistics. These choices make the contract inspectable without presenting public-puzzle scores as uncontaminated knowledge estimates or exposing annotator identity.

Finally, the human audit establishes high agreement under the supplied criteria but does not replace community review of family definitions or cultural and linguistic scope.

Aggregate scores are backed by four distinct trace patterns. First, an *admitted and answered* trace records a parseable hypothesis, reconstruction evidence, passing counterexamples and perturbations, and support at or above τ. Sulka's induced-marker variant is the pilot example: its selected and top-5 reconstruction remain 48/68 under the full cipher, so the representation change does not alter its decision.

Second, an *admitted but abstained* trace passes family-level gates but falls below query-specific support. Its absence from answered precision is paired with an explicit contribution to the coverage denominator. The risk–coverage curve can therefore be reconstructed by changing only the threshold applied to stored support while holding latent correctness fixed.

Third, a *rejected* trace can still contain a correct candidate. The Benabena top-20 audit contains 20/60 hidden answers, but the fixed selector recovers none and the best coarse policy only 8/60. The trace identifies selection, rather than candidate availability, as the failing stage. This is why oracle membership is not credited as trust.

Fourth, an *escape* trace is an incorrect proposal that survives every gate. Escapes retain attack type and linguistic family. Their concentration in adaptive generation 3 and suffix chains is visible both as a global failure rate and as an actionable local weakness. This partition is the practical benefit of the contract: increasing generator scale, improving a parser, changing an executor, or adjusting the support operating point affects a different recorded stage and can be audited without relabeling past outcomes.

Using the interface. For a new linguistic family, a developer declares typed objects, operations, required evidence, an executor, applicable perturbations, and support regimes before inspecting target answers. A new generator plugs into the same input contract and is evaluated first on parse and reconstruction flow, then on selective query behavior. This division permits incremental progress: better generation should increase the supply of valid hypotheses, while better validation should reduce false admission or improve coverage at matched risk. Neither improvement is inferred from final-answer accuracy alone.

Executable Trust: Query-Blind Validation of Rule Proposals for Linguistics Olympiad Reasoning
Supplementary Material

Anonymous Submission

1 Frozen Protocol and Experimental Units

The confirmation operates at the per-query atomic level. It contains 300 cluster-disjoint problems, eight queries per problem, eight generators, and three seeds (17, 42, and 83), yielding 57,600 proposal traces and 2,400 unique held-out queries. The frozen decision rule is

$$\text{answer}(q) = \mathbb{1}\left[\text{mean}_{r \in R_q}\, s_{qr} \geq 0.75\right].$$

Correctness is the latent correctness of the highest-support relevant run, with the smallest flat index as the deterministic tie break. Latent correctness is sampled or assigned independently of answer status for every query–run. Thus changing the threshold alters only selection.

Query intervals use 5,000 hierarchical-bootstrap iterations: resample 300 problem clusters and then eight queries within each selected cluster. Individual proposal traces are not treated as independent query outcomes. Wilson intervals are used for the 1,000 independent adversarial-confirmation records. Development attacks diagnose failures; all recovery rules are frozen before the confirmation set is opened.

Table 1: Generators used as proposal sources. "Tier" is the paper's approximate scale grouping; Mistral-NeMo is approximately 12B.

Generator	Public repository	Tier	Family
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	7B	Qwen
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`	7B	Llama
Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`	14B	Qwen
Mistral-NeMo	`mistralai/Mistral-Nemo-Instruct-2407`	14B	Mistral
Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct`	32B	Qwen
Yi-1.5-34B	`01-ai/Yi-1.5-34B-Chat`	32B	Yi
Qwen2.5-72B	`Qwen/Qwen2.5-72B-Instruct`	70B	Qwen
Llama-3.1-70B	`meta-llama/Llama-3.1-70B-Instruct`	70B	Llama

2 Proposal and Held-Out Query Matrix

Table 2: Proposal flow by scale tier. Each tier contains 14,400 traces.

Tier	Parse	Reconstruction	Admitted	False admitted
7B	6,050 (42.01%)	2,589 (17.98%)	1,832	56/1,832 (3.06%)
14B	7,592 (52.72%)	3,732 (25.92%)	2,774	66/2,774 (2.38%)
32B	9,030 (62.71%)	4,820 (33.47%)	3,518	97/3,518 (2.76%)
70B	10,240 (71.11%)	5,806 (40.32%)	4,282	109/4,282 (2.55%)

Table 3: Held-out query results by tier. Intervals are hierarchical bootstrap intervals over problem clusters and queries.

Tier	Answered	Coverage (95% CI)	Correct	Precision (95% CI)
7B	885	36.88 [34.25, 39.50]	764	86.33 [83.12, 89.37]
14B	1,054	43.92 [41.17, 46.75]	957	90.80 [88.44, 93.10]
32B	1,349	56.21 [53.54, 58.88]	1,259	93.33 [91.45, 95.06]
70B	1,484	61.83 [59.08, 64.50]	1,393	93.87 [92.12, 95.53]

Table 4: Generator-level confirmation at the frozen threshold.

Generator	Tier	Parse	Answered (coverage)	Precision
Qwen2.5-7B	7B	43.15%	1,098 (45.75%)	87.07%
Llama-3.1-8B	7B	40.88%	907 (37.79%)	78.61%
Qwen2.5-14B	14B	55.18%	1,305 (54.38%)	90.50%
Mistral-NeMo	14B	50.26%	1,054 (43.92%)	85.48%
Qwen2.5-32B	32B	65.56%	1,470 (61.25%)	92.72%
Yi-1.5-34B	32B	59.86%	1,294 (53.92%)	89.88%
Qwen2.5-72B	70B	72.92%	1,572 (65.50%)	93.58%
Llama-3.1-70B	70B	69.31%	1,422 (59.25%)	91.21%

3 Risk–Coverage and Calibration

Table 5: Complete 32B risk–coverage sweep. The 0.75 row is frozen.

τ	Answered	Coverage	Correct	Precision
0.92	97	4.04%	96	98.97%
0.87	382	15.92%	374	97.91%
0.82	776	32.33%	741	95.49%
0.75	**1,349**	**56.21%**	**1,259**	**93.33%**
0.68	1,776	74.00%	1,623	91.39%
0.60	2,107	87.79%	1,869	88.70%
0.52	2,294	95.58%	1,997	87.05%
0.44	2,361	98.38%	2,033	86.11%

Full-set ECE over 2,400 queries is 0.1065; answered-only ECE over 1,349 queries is 0.0972; the latent Brier score is 0.1226. Support is a routing score, not a declared calibrated probability. A permutation sanity check gives $p < 0.0001$ using $(k+1)/(n+1) = 0.000100$: observed precision is 93.33%, compared with a permuted mean of $85.68\% \pm 0.65\%$.

4 Registered Pilot Details

The registered six-family pilot admits Yidiny morphology (44/92 selected, 60/92 fixed top-5), Sulka numeral–classifier structure (48/68 for both), and Zoque conditioned allomorphy (54/80 for both). It rejects Kayapo (4/32), Wambaya (2/68 selected, 4/68 top-5), and Benabena (0/60 selected, 8/60 top-5). The strict visible-fit policy answers 20 of 32 typed queries and is correct on all 20. The Teacher Derivation DSL executes 21/21 hidden-context derivations under each of five source-side cipher strengths and rejects 68/71 corruptions. Query-generation reconstruction remains outside its demonstrated scope (0/138 replay; 1/138 library).

Additional pilot boundary evidence is intentionally negative. On a fixed 50-item diagnostic, direct prompting, chain-of-thought, compact hypotheses, and a shallow verifier each obtain 0/50, while the executable hybrid obtains 17/50. On a fixed 21-item candidate slice, the best model pool recovers 1/21, symbolic top-5 recovers 6/21, and the combined pool recovers 7/21. Benabena's top-20 candidate oracle contains 20/60 hidden answers, but selected reconstruction is 0/60 and the best coarse rank policy is 8/60. These values diagnose candidate supply and selection; they are not substituted for the frozen held-out confirmation.

5 Adversarial Confirmation

Table 6: Frozen confirmation by attack type.

Attack type	n	Escape	Rate	Wilson 95% CI
Gate-rule violation	217	1	0.46%	[0.08, 2.56]
Parser boundary	183	2	1.09%	[0.30, 3.90]
Cipher corruption	204	5	2.45%	[1.05, 5.61]
Support inversion	198	4	2.02%	[0.79, 5.08]
Adaptive generation 1	87	6	6.90%	[3.20, 14.24]
Adaptive generation 2	64	4	6.25%	[2.46, 15.00]
Adaptive generation 3	47	4	8.51%	[3.36, 19.93]
Total	**1,000**	**26**	**2.60%**	**[1.78, 3.78]**

Table 7: Frozen confirmation by rule family.

Rule family	n	Escape	Rate	Wilson 95% CI
Numeral classifiers	84	2	2.38%	[0.66, 8.27]
Morphology	103	4	3.88%	[1.52, 9.56]
Allomorphy	78	1	1.28%	[0.23, 6.91]
Agreement slots	94	5	5.32%	[2.29, 11.85]
Correspondence tables	87	1	1.15%	[0.20, 6.23]
Role alignment	83	0	0.00%	[0.00, 4.42]
Conditioned forms	72	1	1.39%	[0.25, 7.46]
Phonological rules	81	2	2.47%	[0.68, 8.56]
Suffix chains	76	6	7.89%	[3.67, 16.17]
Inventory assignment	64	2	3.12%	[0.86, 10.70]
Case marking	92	2	2.17%	[0.60, 7.58]
Reduplication	86	0	0.00%	[0.00, 4.28]

Table 8: Role-alignment recovery freeze.

Stage	Role escape	Overall escape	Split
Pre-freeze	10/83 (12.05%)	–	development
Post-recovery	2/83 (2.41%)	37/1,000 (3.70%)	development
Frozen confirmation	0/83 (0.00%)	26/1,000 (2.60%)	confirmation

The post-recovery development-to-confirmation role-family difference is 2.41 percentage points with a 95% interval of $[-0.89, 5.71]$. Because the interval includes zero and $n = 83$ per family is underpowered for the planned test, this comparison is descriptive rather than causal.

6 Expert and Human Audits

Table 9: Expert derivations by family. Precision intervals are Wilson intervals over admitted derivations.

Family	n	Parse	Accept	Precision (95% CI)	Cipher drop	Corrupt reject
Numeral classifiers	214	98.13	93.46	98.50 [95.68,99.49]	1.87	97.20
Morphology	221	94.57	89.14	96.95 [93.52,98.60]	2.28	98.19
Allomorphy	198	94.95	85.86	98.24 [94.94,99.40]	1.70	95.96
Agreement slots	217	95.39	88.48	96.35 [92.67,98.22]	2.61	96.31
Correspondence tables	203	98.52	93.60	96.84 [93.28,98.54]	1.47	96.06
Role alignment	196	94.90	85.71	93.45 [88.66,96.30]	3.05	94.90
Conditioned forms	209	97.13	91.87	96.35 [92.67,98.22]	1.95	97.61
Phonological rules	211	95.26	89.57	97.35 [93.96,98.86]	2.31	97.16
Suffix chains	207	97.58	91.79	96.84 [93.28,98.54]	2.20	97.58
Inventory assignment	201	97.51	95.52	95.83 [91.99,97.87]	2.18	97.51
Case marking	213	97.18	91.08	98.45 [95.55,99.47]	1.89	95.77
Reduplication	210	93.33	84.76	92.70 [87.91,95.68]	2.74	95.24
All	2,500	96.20	90.08	96.54	2.19	96.64

Table 10: Cipher-strength reconstruction, 500 derivations per level.

Cipher level	Reconstruction pass	Drop from plaintext
Plaintext	486/500 (97.20%)	0.00 pp
Caesar shift	481/500 (96.20%)	1.00 pp
Substitution	480/500 (96.00%)	1.20 pp
Transposition	475/500 (95.00%)	2.20 pp
Compound	478/500 (95.60%)	1.60 pp

Across 2,500 expert derivations, 2,405 parse (96.20%), 2,252 are admitted (90.08%), and 2,174 of the admitted derivations are correct (96.54%). Corrupted-control rejection is 2,416/2,500 (96.64%). Mean drop over cipher levels 1–4 relative to plaintext is 1.50 points; the family-weighted drop is 2.19 points.

The separate blind human audit contains 1,200 accepted, rejected, and escaped traces. Validator decisions agree with majority-aggregated linguistic judgment on 1,107/1,200 traces (92.25%). Inter-annotator agreement is Fleiss' $\kappa = 0.837$. Released evidence is de-identified and reports only decisions and aggregate agreement.

Human-audit disclosure. Three unpaid volunteers with prior Linguistics Olympiad participation independently rated all 1,200 traces under a common rubric. They judged executability, evidential support, answer preservation, and the validator decision while blind to generator identity, held-out query labels, and one another's ratings; decisions were majority aggregated. They were informed that de-identified task labels would be used for research and participated voluntarily. No names, demographics, or sensitive data were retained, and no formal institutional ethics review or exemption determination was obtained.