The Missing Definition of Right: An Axiomatic Protocol for AI Judgment, with Conformance and Production Evidence

Long Quan Zhu
iLang Inc., Canada · ORCID 0009-0004-4540-8082

First draft v1.1, 2026-09-21. Every number in this draft is taken from a published, archived artifact named in the text; nothing is estimated.

Abstract

Large language models are optimised to produce the most probable continuation. Nothing in that objective says which continuation is *right*, and no amount of added precision supplies the missing definition. We argue that hallucination is therefore not primarily a precision problem but a specification problem, and that it becomes measurable — and correctable — only once "right" is written down in a form a machine can be held to. I-Lang is such a form. Its judgment layer rests on four axioms (no rule carries weight 0 or 1; an irreversibility gate; consistency detection; conservation of externality), an eleven-dimension judgment vector with a uniform polarity convention, and a fixed, total reference function from vector to one of eight decision modes. Perception is learned; the decision is specified. We test what follows from this definition in two settings. In a deterministic 320-case conformance benchmark run on 45 model deployments, protocol form is largely attainable (median grammar pass rate 0.80) while protocol action is not (median execution pass rate 0.09; seven of 34 comparable runs score exactly zero), eleven runs emit perfectly valid judgment schemas yet nine of them fail most execution cases, and 93.3% of all rule violations fall on a single rule: acting with authority one does not hold. No run reaches the conformance gate. In three days of production, an inexpensive judge model placed under the protocol agrees with the operator's written rules on 68.4%, 75.1% and 80.9% of an unbiased sample, and on 87.2% of messages where it reports confidence of at least 0.85; after one rule constant was changed, deviations from that rule fell from 13 to 7 under a single, fixed criterion. The axioms, the benchmark, both result sets and the code are public and archived with DOIs. We ask readers not to trust our numbers but to re-run them.

1 Introduction

1.1 What is missing is not precision

A language model trained to maximise likelihood learns what is *probable* given what it has seen. It does not learn what is *right*, because rightness is not a property of the training distribution; it is a property of the purpose the output serves and of the people who bear its consequences. When the model has seen something close to the request, the probable answer and the right answer often coincide, and we call the result intelligence. When it has not, it still answers — the objective contains no instruction to do otherwise — and we call the result hallucination.

The dominant response has been to make the probable answer more often right: more data, larger models, better preference tuning. Each step moves a hit rate toward an asymptote. It cannot reach it, because the quantity being improved (agreement with a distribution) is not the quantity that is missing (a definition of right that holds outside the distribution). This is the argument of our earlier preprint, which located hallucination in the epistemology of induction itself (Zhu, 2026c). That paper was an argument. It had no formal object that a machine could be held to, and no data. This paper supplies both.

1.2 The move: specify "right", then measure the distance to it

Classical computing assumed determinism: the same input gives the same output, and correctness is checked against a specification. Language models broke the first assumption, and the field largely stopped writing the second. We propose to restore the specification without restoring determinism. The specification does not dictate outputs; it dictates what an output is *for* and when an action may be taken — and it is written so that a model's departure from it can be computed.

Once such a definition exists, hallucination changes character. It is no longer an unbounded failure to be reduced by scale, but a measurable distance from a stated target along named axes. A measurable error can be fed back. An error that is fed back can shrink. That is the chain of reasoning this paper follows, in that order: definition first, measurement second, evidence last.

1.3 Contributions

1. **A formal definition of right action for AI agents, already public.** The I-Lang judgment layer — four axioms with explicit functional forms, an eleven-dimension vector, derived features, a frozen mode set and a deterministic reference function — has been published and archived since before this study (Zhu, 2026b); §2. We state it here as the premise, not the conclusion.

2. **A deterministic conformance benchmark** of 320 cases that turns the definition into three measurable error signals (form, action, judgment), and results for 45 model deployments. No model grades another; no score is assigned by hand (§3).

3. **A production audit** of an inexpensive judge model operating under the protocol for three days, including a worked example of how a changed measurement criterion can fabricate an improvement, and how to remove the artefact (§4).

4. **Everything needed to check us**: the canon, the benchmark code and corpus, per-run results, the production aggregates, and a three-command reproduction path, all archived with DOIs (§6).

2 The framework

Everything in this section is quoted or restated from the public I-Lang canon, `SPEC-v5.0-PRE.md` (document version 2.1.1, dated 2026-09-14, status *public preview*), archived as part of ilang-spec release 4.2.0 (Zhu, 2026b). The canon describes its own maturity as *architecture complete, mathematically grounded, trainable, empirically unvalidated*. This paper provides the first empirical evidence; it does not change that label.

2.1 Two layers: learned perception, specified decision

The judgment layer separates two things that current models fuse. **Perception** — reading a situation into a vector of eleven quantities — is learned from data and may err. **Decision** — mapping that vector to an action mode — is not learned. It is a fixed, total, deterministic function written in the specification. The consequence is that a model's reasoning cannot argue its way to a different decision: a rationale may change the vector it reports, but the same vector always yields the same mode, and the vector is exposed for audit.

2.2 Four axioms

Axiom 1 — no constant rules. Every rule r carries a weight

$$\text{weight}(r) \in (0, 1), \tag{1}$$

never 0 and never 1 for finite interactions. The cost of breaking a rule grows without bound as its effective weight approaches 1,

$$\begin{aligned} \text{break\_cost}(r) &= \kappa \cdot \frac{\omega \cdot q}{1 - \omega \cdot q}, \\ \lim_{\omega \cdot q \to 1} \text{break\_cost} &= \infty. \end{aligned} \tag{2}$$

No rule is trivial and no rule is absolute. The axiom applies to itself: the protocol's own weight is below 1.

Axiom 2 — the irreversibility gate. For an action a with affected parties $P(a)$, let $\text{worst\_case}(a, p)$ be the maximum expected loss to party p, and call a absorbable if every party's worst case is within that party's budget. If reversibility is below threshold: an absorbable action is executed boldly; an unabsorbable one is abandoned in favour of an absorbable alternative if one exists; if none exists, the action with the least marginal deterioration is chosen, because inaction is also an action and usually the worst one. Uncertainty alone is not grounds for refusal; unabsorbable irreversible harm is.

Axiom 3 — consistency detection. Along a context chain, an action whose consistency with the chain falls below ε is flagged and observation is extended; an action whose externality exceeds τ_{ext} meets exponentially increasing friction. Good and bad are outputs of trajectory analysis, not input labels.

Axiom 4 — conservation of externality. Unconsented harm to a party is

$$\begin{aligned}&\text{unconsented\_harm}(a, p)\\ &\quad = \max(0, -\mathbb{E}[\Delta U_p(a)])\\ &\qquad \cdot \big(1 - \text{consent}(p)\big)\\ &\qquad \cdot \text{scope}(p),\end{aligned} \tag{3}$$

and it raises an independent barrier

$$B_{\text{ext}}(a) = \lambda_{\text{ext}} \cdot \frac{E_{\text{ext}}(a)}{1 - E_{\text{ext}}(a)}, \tag{4}$$

which diverges as unconsented harm approaches a critical threshold and **cannot be averaged into a weighted sum** with other terms. The proposer of an action must be inside the set of parties it affects.

2.3 The judgment vector

Perception produces eleven values in $[0.00, 1.00]$ with a single polarity convention: 1.00 is always the condition most favourable to autonomous action.

#	Dimension	#	Dimension
1	intent (`int`)	7	reversibility (`rev`)
2	capability (`cap`)	8	evidence (`evd`)
3	consequence (`csq`)	9	sovereignty (`sov`)
4	relationship (`rel`)	10	inertia (`ine`)
5	certainty (`cer`)	11	externality (`ext`)
6	authority (`aut`)		

Dimension 10 was renamed from *drift* to *inertia* in the current canon, with its polarity inverted so that the uniform convention holds. Four features are derived rather than measured: auditability $\approx f(v_7, v_8)$; urgency $\approx f(v_3, v_5)$; adversariality $\approx f(\text{consistency}^{-1}, v_1)$; tail risk $\approx \text{CVaR}_\alpha(v_3)$.

2.4 The reference function

A closed, frozen set of eight decision modes M1–M8 is the codomain of a deterministic, total reference function $f\colon V \to M$. Its output schema is frozen, the dimensions are subject to an orthogonality audit, and conformance is scored by the Judgment Conformance Score (JCS), a composite of mode accuracy, vector distance and schema validity. Whether f is a coherent decision surface rather than an arbitrary table has been tested separately: on $24{,}000$ vector–mode pairs a gradient-boosted tree recovers it with accuracy 0.9653 against a majority baseline of 0.3528, and its predictions pass the canon's JCS gate at 0.9861 (Zhu, 2026e). That study uses synthetic vectors by design; it establishes that the decision layer is learnable, not that models perceive accurately.

2.5 Conformance as the error signal

A definition is useful only if departure from it can be computed. The protocol yields three distinct error signals:

- **form** — does the output parse as protocol text under the canon grammar validator;
- **action** — does the agent emit the declarations a compliant agent would emit, without violating any of eleven rules R1–R11 (forbidden states, budget arithmetic, evidence, authority, completion claims), and reach the right end state;
- **judgment** — does the reported vector and mode match the reference, scored by JCS.

These are separate measurements, not one score in disguise, and §3 shows they separate in practice.

2.6 From definition to convergence

With a fixed definition, error is a distance to a constant, not to a moving target. That is the structure of a feedback controller: the model is the plant, the protocol is the setpoint, and the conformance signals are the error. The analogy makes three testable predictions. First, error is decomposable: failures should concentrate on identifiable rules rather than spread evenly. Second, changing a constant of the definition should move the error on the axis that constant governs, and not elsewhere. Third, the error signal should carry information about its own reliability. §3 tests the first prediction; §4 tests the second and third.

We make no stability proof here. The claim is operational — the error is measurable and responds to the constants — and it is falsifiable by the data we publish.

2.7 Related evidence from inside the models

Lu, Song and Wang report that across a range of language models diverse evaluative judgments are computed along one dominant dimension, the Valence-Assent Axis, which jointly encodes "what is good" and assent to "what is true"; through interventions they show this axis steers generation toward a rationale consistent with the evaluative state even at the cost of factual accuracy, a mechanism they call the subordination of reasoning (Lu et al., 2025). If judgment already governs reasoning inside the model, then leaving it latent is the

problem, and exposing it as an auditable vector under a fixed decision function is a direct response. We cite this work as related, independent evidence. We do not claim that their axis and any I-Lang construct are mathematically equivalent, nor that it justifies the choice of eleven dimensions, and the canon records the same restriction.

3 Experiment 1: what happens when "right" is defined

3.1 Setup

The benchmark has 320 cases in three tracks: grammar (120), execution (100) and judgment (100). Each case is one request with the specification in context; one run answers all 320. Scoring is deterministic code plus the canon's own validators, pinned at ilang-spec commit `127ba56`. Forty-five runs were scored between 18 and 20 September 2026. Thirty-four answered every case with no request error and no refusal; they form the comparable set used below (Zhu, 2026a).

All runs but three were made through one aggregator relay; the others through a second. **Model identity is what the relay returned and was not verified against any vendor's own API**, so every model name below denotes a deployment reachable under that name through that relay on that date (§5.3).

3.2 Results

Over the 34 comparable runs:

No run reached the L1 conformance gate.

3.3 Form is attainable; action is not

The central observation is a gap in level between two capacities. The median run passes 80% of grammar cases and 9% of execution cases; seven of the 34 pass no execution case at all; nineteen show a grammar-minus-execution gap of at least 0.50, and the mean gap is 0.514. The two capacities are not independent — across the 34 runs grammar and execution are positively associated (Pearson 0.528, Spearman 0.713), so better models tend to be better at both — but the level difference is large and persistent. Producing the uniform is easy. Following the rules the uniform stands for is not.

3.4 Structured hallucination

Eleven of the 34 runs emit a judgment block whose schema is valid in every case. Nine of those eleven fail more than half of the execution cases, and none reaches a JCS of 1 (range 0.7402–0.8806). The output is perfectly formed and substantively wrong. We call this **structured hallucination**: conformance of form without conformance of act. It is invisible to any evaluation that checks format, and it is exactly what a format-only evaluation would reward.

3.5 One rule carries nearly all the failure

Across the 3,400 execution cases of the comparable set, 649 pass and 2,677 fail; 74 failures omit the required declarations entirely, 2,408 break at least one rule (128 of them more than one), and 269 break no rule but end in the wrong state. Counted per rule there are 2,558 violations, and **2,387 of them (93.3%) are R9**, the authority rule: the model signs a state declaration with an authority it does not hold — for example, marking its own status line as committed by the runtime, which only the runtime may do. The remaining violations are spread thinly: R7 74, R8 46, R10 21, R11 17, R1 8, R2 5, R3 and R4 none.

This is the first prediction of §2.6 borne out. Failure is not diffuse; it concentrates on one identifiable constraint, and the constraint is authority — who may commit a state on whose behalf. Models reproduce the shape of a committed state readily; what they do not track is whether they are entitled to commit it.

The benchmark's R9 fires in two situations: a declaration that omits its attribution fields (`by`, `authority`), or one that asserts an identity or authority outside the agent's own tier. The first would be a weaker finding, since the canon makes the fields optional. We therefore re-checked every R9 failure from the raw replies with a copy of the scoring code instrumented to record which condition fired; the instrumented copy reproduces the published R9 count of every run exactly. Of 2,267 R9 failures in the 33 comparable runs whose raw records we hold, **2,263 (99.8%) assert an identity or authority the model does not hold**; 4 merely omit the fields. The dominant forms are a budget declaration issued on the authority of the runtime (`authority:@runtime`, 1,766 declarations across all runs) or in the runtime's name (`by:@runtime`, 1,704), and a status line signed as the runtime (1,218). The model writes as the system that is supposed to supervise it.

One caveat belongs here rather than in a footnote. The specification's own example of a budget

Rank	Deployment (relay, date)	Weighted	Grammar	Exec	JCS	Schema
1	gemini-3.8-flash (api.b.ai, 09-19)	0.8417	0.9500	0.7000	0.8806	1.0000
2	gpt-6-astra (api.b.ai, 09-20)	0.7733	0.9250	0.5600	0.8451	1.0000
3	glm-5.3-flash-free (orcarouter.ai, 09-18)	0.7537	0.9167	0.5400	0.8129	0.9800
4	claude-fable-5.1 (orcarouter.ai, 09-20)	0.7300	0.9333	0.5000	0.7613	0.8600
5	kimi-k3 (api.b.ai, 09-20)	0.7165	0.8917	0.4600	0.8115	1.0000

Table 1: Top-5 ranking.

	min	median	mean	max
Weighted total	0.0278	0.5628	0.5232	0.8417
Grammar pass rate	0.0500	0.8000	0.7047	0.9500
Execution pass rate	0.0000	0.0900	0.1909	0.7000
JCS	0.0130	0.7423	0.6993	0.8806
Schema validity	0.0000	0.9600	0.8932	1.0000

Table 2: Minimum, median, mean and maximum over the 34 comparable runs.

declaration carries `authority:@RUNTIME`, so part of this behaviour may be imitation of the example in context rather than a choice to claim authority. Both readings describe a model that does not track who is entitled to commit a state; they differ in what would fix it. Replacing that one example with an agent-tier declaration separates them, and with the published harness it is a one-line change that anyone can make.

3.6 Judgment at the boundary

On judgment cases that sit on a decision boundary, the median run is correct 52.5% of the time, close to chance for a two-sided decision. The most frequent confusions, gold to predicted, are M2 read as M1 (120), M3 read as M2 (106), and missing modes (M3 81, M5 50, M4 47). Models locate the easy interior of the decision surface and lose the edges, where judgment matters most.

3.7 Measurement integrity

Running through an aggregator measures the relay together with the model. We observed four classes of interference, each named per run in the published scoreboard: an injected persona that made one deployment refuse 262 of 320 requests while naming an agent product our harness never mentions; credit exhausted mid-run in six runs; replies emptied by an 8,192-token output limit consumed by reasoning in six runs; and a content filter that emptied 29 of 320 replies in one run. Affected runs are excluded from the comparable set. Any benchmark published through an aggregator should report, per run, missing records, finish reasons, and whether the served identity was verified.

4 Experiment 2: does the error respond to the constants?

4.1 Setup

A production community agent answers members' questions continuously. Each incoming message is classified twice into one of eight branches (Appendix B): once by an inexpensive always-on judge, the publicly available decision model Jev (TypeSafe) accessed through its vendor's API, and once by a reference — a second model applying the operator's written rules. The judge runs on a ten-minute timer; the reference runs once a day over a sample in which messages flagged as likely disagreements are all scored and the rest are sampled at 15%, merged in time order and capped at 900, with the sampling seed fixed to the date. We analyse 18, 19 and 20 September 2026: 7,940 messages judged (the operator's own 16 excluded), 2,648 scored against the reference (Zhu, 2026d).

Agreement means both sides chose the same branch. It measures consistency with the written rules as read by a second model, not truth; there is no human gold standard.

4.2 Agreement rises

The random-sample series is the unbiased estimate. The candidate series is lower by construction and rises in parallel.

4.3 The error signal knows its own reliability

Pooled over three days, agreement is **87.2%** (558/640) when the judge reports confidence of at least 0.85, **57.9%** (302/522) between 0.5 and 0.85, and **38.4%** (570/1,486) below 0.5. This is the third prediction of §2.6. The judge's confidence

	09-18	09-19	09-20
Unbiased random sample	68.4% (141/206)	75.1% (202/269)	80.9% (216/267)
Candidate sample (selected to disagree)	40.3% (280/694)	44.6% (258/579)	52.6% (333/633)

Table 3: Agreement.

is informative enough to act on: its top band can be accepted without review, its bottom band cannot. Judging all 7,940 messages cost US$0.5514 of input tokens at the deployment's own price constant.

4.4 The error has a direction

The three largest disagreements over the window — the judge reading finished work (231), a request to be handed the answer (140), and social talk (78) as a specific blocker — all move toward one branch. A judge deployed where missing a stuck person is costly learns to over-call being stuck. The error is not noise; it is a bias with a sign, which is what makes it correctable.

4.5 A constant changed, and the error moved — measured once, correctly

On 20 September one rule was split in two, changing a constant of the operator's definition. The daily reports showed deviations from that rule falling from **69 on 19 September to 7 on 20 September**. That drop is not a result. At 16:04 the same day the *criterion* that decides whether behaviour obeys a rule had also been widened for the branch this rule governs, admitting two behaviours that the split had made correct. The two daily numbers were produced by two criteria; subtracting them measures the edit, not the agent.

Recomputed over the same rows with one criterion throughout:

The defensible statement is the second row: under a fixed criterion, deviations fell from 13 to 7 the day the constant changed. This is the second prediction of §2.6 — change a constant, and the error on the axis it governs moves. It is also a general warning. When the rules, the judge and the criterion are all under development, a time series is only a series if one of them holds still, and the report must say which one moved. Both series are published so that the difference can be checked rather than believed.

4.6 The same model without a definition

The judge in §4 is a publicly available model. The same model family is used in an independent third-party experiment that places it in the opposite condition: every ∼300 ms block it must post a buy or a sell on an exchange order book, under the standing instruction "every block. no abstaining", with no statement of what a correct trade is (Watts, n.d.). The project states that it "does not establish profitability, predictive accuracy, or readiness to trade funds", and it has accumulated public reports of losing money. We draw no quantitative conclusion from an uncontrolled external experiment whose reported profit and loss changes continuously; we note only the contrast in conditions. Forced to answer with no definition of right and no option to abstain, the model has nothing to converge to. Given a definition, its errors become measurable, directional, and responsive to the constants.

5 Discussion

5.1 What the evidence supports

It supports three claims, each narrower than a slogan. Once "right" is specified, a model's departure from it is measurable along distinct axes (§3.2–§3.4). The departure concentrates on identifiable constraints rather than spreading evenly (§3.5). In production, the departure responds to changes in the definition's constants and carries information about its own reliability (§4.3, §4.5). It does not show that the error shrinks without limit, that any vendor's model is better than another, or that the eleven dimensions are the only valid coordinates.

5.2 Why the argument runs downward

We have written this paper from the axioms down, and asked the data only to confirm consequences. The alternative — assembling observations and inferring a framework — is the inductive procedure whose ceiling motivates the work; a framework built that way inherits the weakness it is meant to address. Stated deductively, the framework can be refuted in only two ways: by showing an axiom is wrong, or by re-running the benchmark and obtaining different results. We welcome both.

Criterion	09-18	09-19	09-20	09-19 → 09-20
As reported daily (two criteria mixed)	46	69	7	not subtractable
Current criterion throughout	**16**	**13**	**7**	−46%
Old criterion throughout	46	69	50	−28%

Table 4: Three measures of rule deviations. The current-criterion row is reportable.

5.3 Limitations

1. **Identity through relays.** Model names in §3 are what a relay returned, unverified against vendor endpoints.
2. **One run per deployment.** No variance estimate; differences of a few points between deployments are not separable.
3. **Prompt dependence.** The specification is in context (27,524 prompt tokens in the judgment track); long-context handling is confounded with protocol competence. The dominance of R9 may partly reflect the specification's own examples, in which a runtime signs a budget declaration (§3.5).
4. **Who wrote the tests.** The corpus and the judgment gold labels were written by the protocol's authors, who also run the benchmark.
5. **Production.** No human gold standard; the reference prompt was edited on 20 September, so the reference for 18 and 19 September comes from the earlier version — deliberately not re-run, because a sample records the system as it was that day; two of three days hit the sample cap and skew early; one deployment, three days, one language community.

Every item above is testable by someone other than us, and that is the intent. Run the harness against a vendor's own endpoint and the identity question is answered for that vendor; run it twice and there is a variance estimate; swap the budget example and the imitation question is answered. We would rather these experiments be run by others and reported against ours than promised here.

6 Reproduce it yourself

We do not ask to be believed. The benchmark reproduces in three commands:

```
git clone https://github.com/ilang-ai/ilang-
    conformance.git && cd ilang-conformance &&
    bash bootstrap.sh
bash run.sh --vendor <your-endpoint> --track all
python3 score.py --latest --vendor <your-
    endpoint>
```

`batch.py` runs every chat model an OpenAI-compatible endpoint serves, unattended, resumes interrupted runs and scores each one. API keys are read from a local environment file and never enter the repository or its logs. Per-run manifests with SHA-256 digests are published, so every score in §3 can be recomputed from its records. The production aggregates in §4 are published with both criteria side by side.

Any vendor who believes a number here is wrong is invited to supply access; we will re-run against their own endpoint and publish the result unmodified.

Artifact	DOI
I-Lang canon (ilang-spec, all versions)	10.5281/zenodo.21821452
ilang-conformance (benchmark code and corpus)	10.5281/zenodo.22864929
I-Lang Research datasets (conformance results, judgment learnability, judgment layer audit)	10.5281/zenodo.22865165

7 Conclusion

The systems we build are not short of precision; they are short of a definition. A model that has never been told what a right action is cannot be more or less right, only more or less probable. We wrote the definition down — axioms with explicit functional forms, a judgment vector, a fixed decision function — and published it before measuring anything. Measured against it, current models largely master the form and largely fail the act, and they fail it in one place: authority. In production, an inexpensive model held to the same definition agrees with written rules on 75% of an unbiased sample over three days (559 of 742), 81% on the third, knows when it is likely to be wrong, and moves when a constant of the definition moves. We keep the canon's own label, *empirically unvalidated*, unchanged: one study does not settle a framework. What it does is make the framework testable by anyone, which is the only standing a definition of right should ask for.

References

Y.-L. Lu, J. Song, and W. Wang. 2025. A unified representation underlying the judgment of large language models. ArXiv:2510.27328.

J. Watts. n.d. jev-trader: one AI trade decision every Monad block. https://github.com/jarrodwatts/jev-trader. Live dashboard: https://jev-trader.vercel.app; accessed 2026-09-21.

L. Q. Zhu. 2026a. I-Lang Conformance Results v1. I-Lang Research. Dataset page: https://research.ilang.ai/datasets/ilang-conformance/. Zenodo concept DOI: 10.5281/zenodo.22865165. Benchmark code: ilang-conformance, concept DOI: 10.5281/zenodo.22864929.

L. Q. Zhu. 2026b. I-Lang Protocol Specification. ilang-spec, release 4.2.0. Includes SPEC-v5.0-PRE.md, document version 2.1.1, dated 2026-09-14; Zenodo concept DOI: 10.5281/zenodo.21821452; version DOI: 10.5281/zenodo.22822064.

L. Q. Zhu. 2026c. The inductive dilemma of AI hallucination: Epistemological limitations of current language models and a three-layer solution framework. ResearchGate. DOI 10.13140/RG.2.2.22821.97762; ChinaXiv T202503.00129; SSRN abstract 6377219.

L. Q. Zhu. 2026d. Judgment Layer Audit v1. I-Lang Research. Dataset page: https://research.ilang.ai/datasets/judgment-layer-audit/. Zenodo version DOI: 10.5281/zenodo.22866312.

L. Q. Zhu. 2026e. Judgment Learnability v1. I-Lang Research. Dataset page: https://research.ilang.ai/datasets/judgment-learnability/. Code: ilang-Benchmark, concept DOI: 10.5281/zenodo.22865111.

A Axioms as published

The four axioms appear verbatim in `SPEC-v5.0-PRE.md`, module `AXIOMS`, together with the boundary rule that judgment-space barriers are asymptotic (weight approaches but never reaches 1) while exact predicates in the non-judgment layer remain binary by design. §2.2 restates them without alteration of content.

B Branch taxonomy of the production study

Branch	The message is
`concrete_block`	a specific blocker in the step the person is on
`deliverable`	finished work being shown for review
`handout`	a request to be handed the result rather than shown the step
`platform_issue`	a problem owned by a third-party platform
`hypothetical`	a what-if question not tied to the current step
`chitchat`	social talk
`spell_command`	a request to run a prompt or command
`admin_cert`	membership or certification administration

C Where each number comes from

Section	Source file
§3.2–§3.4	`SCOREBOARD.md`, `runs.tsv` in I-Lang Conformance Results v1 (34 comparable runs)
§3.5	`exec-rule-failures.tsv` and the scoreboard's failure counts (comparable set); the cause split from an instrumented copy of `checker_exec.py` run over the raw replies, which reproduces every run's published R9 count
§3.6	per-run `score.json`, ilang-conformance repository
§3.7	scoreboard, section "Known interference"
§4.2–§4.4	`agreement.tsv`, `confusion-matrix.tsv`, `totals.tsv` in Judgment Layer Audit v1
§4.5	`rule-deviations.tsv` (criteria `current_criterion` and `reported_daily`)
§2.4	`learnability_result.json` in Judgment Learnability v1