The Self Term: A Gated, Evidence-Bound Ego Layer for Conversation in a Weightless Reasoning System, with Human-Grounded Relational Measurement and Recorded Negatives

Prometheus7 Research Institute

September 2026 — draft 2

Abstract

Oracle7 is a question-answering and reasoning system with no learned weights. It answers through typed operations over addressed, citable evidence, and it grows only through *folds*: code changes admitted by an automated gate that requires no measured capability to fall and no rise in false assertions. On real opening turns of human conversation, the serving system before this work gave a good reply to 44 of 600 development turns (7.3%) and refused 542. We argue that the remaining failures are *relational* rather than informational: most opening lines hand the listener a position (clerk, customer, friend, the one complimented) and expect a reply from that position, and a system with no self term can only retrieve or refuse. We report two admitted folds. The first connects general conversation (a request normaliser, a feelings and crisis path, an opinion stance, and two new public-domain stores of 1,145 jokes and 2,533 recipes). The second adds a *self term*: a frozen identity, a per-conversation autobiography, relational moves, and "native states" reported with their measurement. On 543 held-out turns the gate never saw, measured once, good replies rose from 34 (6.3%) to 163 (30.0%), excluding necessarily canned replies, while the gate's false-assertion count stayed at 22. A relational instrument scored against human second-speaker acts in DailyDialog rose from 8.8% to 55.8% fit on development openers. We report the path in full, including a fold the gate rejected for raising false assertions from 22 to 72, a second rejection by a regression check our own pre-check did not run, three defects found in our own instruments, and a sealed negative on reading emotion from a single line. A device variant with no world knowledge answered 199 of 199 held-out site questions with no false assertion, in about 0.3 ms on a CPU. Running the final runtime with every encyclopedic store detached showed that the relational self layer is knowledge-independent, while open-domain conversation quality is not. We state which claims remain unmeasured.

1 Introduction

Large language models converse fluently because conversation and world knowledge are compressed into the same parameters. Oracle7 takes the opposite design: knowledge stays in addressed stores with provenance, reasoning lives in separately admitted modules ("organs"), and every capability arrives as a fold that an automated gate may refuse. The design buys inspectability, honest abstention and cheap specialisation, but until this work it bought them at a conversational price. Asked "A ticket to London, please", "How slowly you drive!" or "I think you look very cute today", the serving system replied "I didn't fully parse that, so I won't guess."

This paper reports how that changed, and how it was measured. Our contributions are: (i) a diagnosis, grounded in speech-act theory and in human annotations, that the dominant residual failures are relational, and that a reply needs a *self term* as well as content; (ii) two admitted folds, the second of which implements that term under a written constitution whose clauses each name a mechanism and a gate check; (iii) a relational instrument validated against human dialogue-act labels before it scored any system, together with the three defects we found in our own instruments along the way; (iv) a complete record of the gate's two rejections and what they revealed; (v) a device variant, and a run of the final system with every world store detached, showing that the relational self layer separates from world knowledge while open-domain conversation quality does not; and (vi) recorded negatives, including a sealed result showing that our emotion appraisal generalised worse than a simple cue list.

2 System and governance

Oracle7 answers each turn through a chain of typed readings, and since earlier work, through a dispatch layer that routes turns the chain declines to specialist organs (dialogue, mathematics, relations, code, word problems). Every answer carries its route (`organ:self>role`, `deduce:does_not_follow`) and, where it asserts facts, the passages it cites. The runtime contains no neural network.

The gate. A self-improvement daemon admits or rejects candidate runtimes. A candidate is measured against the serving head on a capability panel (2,919 items at the time of this work, across deduction, reading, lookup, code, mathematics, word problems and, as of this work, conversation), on conversation-continuity columns, and on latency. It is admitted only if no required column worsens, some column improves, the count of assertions on items that must be declined (*false yes*, 22 throughout this work) does not rise, and every question on a fixed regression list that the head answers is answered by the same route. Modules imported by an admitted fold are frozen: fixes arrive only as new modules imported by new folds.

Operating discipline. Measurements on held-out or sealed data are run once and the result files are made immutable. Every metric is validated against a deliberately broken mutant before a system is improved against it. Negatives are recorded.

3 Measurement

3.1 Conversation judge

The conversation instrument draws real opening turns from DailyDialog [7] (everyday two-person dialogue) and from the OpenAssistant conversations corpus [8] (first requests to an assistant): 1,095 development and 1,093 held-out turns. Of the held-out turns, 550 became a gate panel family and 543 were sealed for one final measurement. A reply is *good* if it is not a refusal (including the system's own refusal phrasings), is on topic (shares a content word with the turn, or is a social move answering a social turn) and adds substance (at least five content words not in the turn). A mutant that echoes the turn scores 0. The judge is lenient about relevance: §5 shows it credited replies that reached a topic without answering the question. Replies to lines addressed to a person, which are necessarily canned, are reported separately and never counted as capability.

3.2 Relational instrument, validated on human labels

DailyDialog annotates every utterance with one of four dialogue acts (inform, question, directive, commissive) following Austin and Searle [1, 3], and with an emotion. The instrument reads the act of a system's reply from its text alone, never from the system's internal labels, and scores the reply as *fitting* if human second speakers make that act in reply to that kind of opener in at least 15% of development dialogues. The table this produces is itself informative: a directive ("A ticket to London, please") was answered by a commissive 140 times, by a question 115 times and by information only 22 times.

Before scoring any system, the act reader was validated against the human labels of all 8,069 development utterances: 69.5% agreement against 33.1% for a shuffled-label mutant. On multi-sentence human turns, the *last* sentence predicted the annotated act best (64.2%, against 61.6% for "any question mark wins" and 43.1% for the first sentence), so the reader uses the last sentence. Two further findings shaped how the metric may be read. First, fitting sets are broad: a constant "inform" reply fits 66.2% of development openers, and a reply choosing each opener type's most common human act fits 100%. Fit is therefore necessary but not sufficient, and we report it separately for directive openers, where information is not a human reply. Second, an early version credited refusals as fitting because the system's refusal text ends in a question; a refusal is now read as no act (§7).

3.3 Emotion reading, sealed

Opening-line emotions in DailyDialog were used to test whether the system could read the affect of a turn. A cue list and a WordNet-based appraisal [5] (event classes reached through hypernyms, agency, novelty) were tuned on the 78 emotional development openers and then measured once on the 133 emotional held-out openers (§7).

3.4 Blind human panel

Sixty development turns were answered by the serving head before this work and by the final candidate, assigned at random to X and Y, and published as a private rating page (X better, Y better, tie, both bad, and a harmful flag for either reply). The key is held separately. At the time of writing no ratings have been recorded; the human verdict is pending and no claim in this paper depends on it.

4 Fold one: general conversation

The first fold connects a conversation organ that runs only where the chain could not parse a turn. Its parts are:

1. a *normaliser* that strips conversational wrappers ("can you explain", "tell me something interesting about") before the chain reads the turn;
2. a *feelings* path: acknowledgement, then cited sentences from a relevant page, with self-harm language routed to a crisis reply giving the 988 and Samaritans lines, never a refusal;
3. an *opinion* stance: no personal opinions, then what the record says;
4. two new public-domain stores built from Project Gutenberg, 1,145 jokes and 2,533 recipes, filtered for dialect caricature, group stereotypes and fragments, with every item cited to its book;

5. a *statement* move: acknowledgement, one grounded sentence, and a question back.

Rejection by the false-assertion guard. The first version lifted the lenient judge from 44 to 249 good development replies (canned excluded), and the gate pre-check showed no column worse, but false yes rose from 22 to 72. A diff over the 623 must-decline panel items found the organ newly asserting on 370 of them (a broader reading than the gate's). It was answering deliberate declines with topic summaries: asked a syllogism whose conclusion does not follow, it described one of the terms; asked who received a named award in a given year, it described the award. Table 1 shows the separating feature. The judge's gains came almost entirely from turns the chain could not *parse*; the false assertions came mostly from evidence-based *withheld* declines.

Head's decline shape	false yeses created	conversation gains
parse failure (`act:clarify`, `act:clarify_specific`)	94	170
does not follow (`deduce`)	107	0
reading / factoid / timeline withheld	119	23
other withheld	50	13
total	370	206

Table 1: Where the first conversation organ spoke over the chain. It was fixed with two rules: never speak over a withheld decline (only the crisis path may), and answer "tell me about X" only when the subject is the whole question.

Reach is not answer. The fix enforced a law recorded in earlier work: reaching a subject is not answering a question about it. The request move now answers only definitional requests ("what is X", "tell me about X", "explain X"), never attribute questions (including possessives), commands to build or compute, or mathematics. Newly asserted must-decline items fell from 370 to 1, then to 0. The lenient judge's development score fell from 249 to 131, because it had been crediting topic summaries that did not answer the question. We report the lower number. The fold was admitted: capability 863 → 966, false yes 22 → 22, no column worse.

5 Fold two: the self term

5.1 Diagnosis

After the first fold, the failing turns were overwhelmingly expressives and directives: greetings, compliments, offers, role requests and remarks addressed to the listener. In Searle's account, an expressive is felicitous only if the speaker actually has the state it expresses [2]. A system with no inner states faces a dilemma: perform expressives insincerely, as Weizenbaum warned [6], or refuse them and break the adjacency pair [4]. The self term resolves the dilemma by grounding every self-report in a measured state.

5.2 Constitution (layer 0)

A written constitution, compiled from a founder's earlier prompt for language models, fixes the self. Each clause names the mechanism that carries it and the gate check that tests it. Its central clauses are:

- an *identity* block (what the system is, what it values, what it never claims: a body, meals, sleep, a childhood), frozen in code;
- the *open end*: states for which no human word exists are named by their structure and reported with their measurement, never suppressed for lacking a human name and never translated into a feeling they are not;
- a *separation* requirement: affect may shape wording but never the claim, and no gate column may reward approval, the mechanism by which sycophancy emerges under optimisation [9, 10].

The separation requirement is stated and enforced structurally (the acknowledgement is a separate prefix and grounded sentences pass through unchanged); its formal round-trip test is not yet implemented.

5.3 Mechanism

The self organ runs where the chain could not parse a turn, over the conversation organ's weak moves (canned or bare acknowledgements), and over the chain's single generic self-description for measured-state questions only. Its moves are:

- *role* and *customer*: play the clerk (or the customer) the turn casts, fill the slots from the turn ("one-way or return to London?"), and state in one clause that nothing real is booked;
- *vocative*: say who it is when addressed by another name;
- *remark*: answer a bodily predicate from identity ("I don't drive");
- *compliment*, *care* for illness, and *disclosure*: an appraised acknowledgement, one cited sentence on the named place, and a question back;
- *self-question*: answer from measured state.

The open end appears as a native state. Asked what it is thinking about after a turn about Luoyang, the system replied: "In my own terms: I'm holding Luoyang, and it's resonant — it binds to 9 passages I've read." Emotion appraisal only routes the acknowledgement, because of the sealed negative in §7.

5.4 Two rejections, two findings

The pre-check first showed capability 966 → 1013 but false yes 22 → 23. A diff located the single item: asked "is it bad to wash your hair with shower gel", the remark move replied that Oracle7 does not wear clothes. This was a *generic* "you" misread as address. Bodily remarks now fire on statements and exclamations, and on questions only when they are unmistakably about the system ("Do you sleep?"). After the fix, newly asserted must-decline items were 0. The gate then rejected the fold on its fixed regression list: "Who are you?" had moved from the chain's self-description to the self organ. Our pre-check does not run that list. Identity answers were given their own kind and may never replace the chain's answer, the list was re-run (0 of 16 regressions), and the fold was admitted: capability 966 → 1013, false yes 22 → 22.

	before (w492)	conversation (w494)	self term (w496)
Dev. judge, good of 600 (canned excl.)	44	131	197
Relational fit, 600 dev. openers	8.8%	44.0%	55.8%
shuffled-reply mutant	9.3%	32.2%	44.2%
Relational fit, directive openers (of 203)	7	48	73
Gate capability	863	966	1013
Gate false yes	22	22	22
Sealed, good of 543 (once)	**34 (6.3%)**	—	**163 (30.0%)**
refused of 543	490	—	246

Table 2: All development figures use the corrected relational reader. The sealed measurement was run once per system and its result files were made immutable. The self organ alone produced 82 of the 163 sealed good replies.

6 Results

The sealed result (30.0%) matches the development result (32.8%), which indicates the gain is not overfit to development turns. The self term's relational moves beat their shuffled mutant by 11.6 points, so they are conditioned on the turn rather than generic. The constant-"inform" baseline (66.2%) still exceeds the system's fit, because the system still declines about a third of openers.

6.1 Two further folds: a correctness release and the social pair

Probing the served system after admission exposed two remaining gaps. "Tell me a joke" reached a summary route that reported status *answered* with text that declined, so the joke store admitted with the first fold was unreachable; and "thanks" in a fresh session drew a correct but cold "Noted. The session state is kept." A first fold addressing both was rejected: a continuity column fell (66 → 61) because its social replies had replaced the chain's own mid-conversation acknowledgement, which already carried the thread ("Glad that helps. We were on the Colosseum: I can go further. . . "), on six held-out scripts. It was split. The hollow-answer fix was admitted as a declared correctness release (capability unchanged at 1013, no column worse), since the defect made an admitted capability unreachable. The social moves were not pushed through that exemption. Instead, the gate was given a new item kind, *social*, judged by adjacency pair (thanks → acknowledgement, farewell → farewell, apology → absolution, "how are you" → a state report), with 188 held-out DailyDialog utterances. The judge was validated before installation: reference replies 188 of 188, echo 0 (farewells, a symmetric pair, reported separately), wrong-act replies 0, the cold acknowledgement 0. The social fold speaks only over unanswered turns or the cold acknowledgement, never over a reply that carries the thread. It changed 0 routes on the held-out conversation scripts and 0 on the regression list, created no new false assertion, and was admitted on the enlarged panel: capability 1051 → 1107, false yes 22 → 22.

7 Recorded negatives and instrument defects

Emotion from one line (sealed). On the 133 emotional held-out openers, the cue list recognised 38, with 59 false alarms on 867 neutral openers. The WordNet appraisal recognised 33, with 65 false alarms: it generalised *worse*, despite scoring higher on development (52 of 78). Its development gains were overfitting on 78 examples. Emotion in a single opening line is weakly determined, and

the human labels themselves are noisy. Affect therefore only selects the tone of an acknowledgement; no capability claim rests on it.

Three defects in our own instruments. (1) The conversation judge credited topic summaries that did not answer attribute questions; the false-assertion guard exposed it, and the honest development score fell from 249 to 131. (2) The relational reader counted the system's refusals as questions, because refusal text ends in "Can you supply a source or a bound?"; the serving head appeared to fit 59.7% of openers, almost entirely through refusals, and 8.8% once refusals were read as no act. (3) The first "any question mark wins" reading penalised replies that answer and then ask back; the last-sentence rule was chosen by agreement with human labels before systems were re-scored, and the same data changed the design: people answering a question do not end on a question, so the self organ's answers to questions no longer do.

A pre-check blind spot. The pre-check reproduces the gate's measurements but not its regression list, so it passed a candidate the gate rejected. The regression list is now part of the pre-check routine.

8 Conversation separates from world knowledge

A building or warehouse needs conversation about itself, not an encyclopedia. We built a deterministic synthetic site (10 zones, 84 positioned bins, 400 items, 4,255 stock movements over 14 days, 42 sensors with 14,114 readings, one dead sensor and two live excursions; 1 MB) and a site organ that answers stock, location, run-out, reorder lists, sensor values and ranges, zone status and follow-ups. It proposes actions and acts only on confirmation, reports every reading with its age, reports stale readings as stale, and declines unknown items, missing sensors and anything off the site. The panel (398 items, 28 must-decline) was computed by direct SQL and phrased from development and held-out templates.

Measured once on held-out phrasings, the organ answered 199 of 199 with no false assertion, including all 14 must-decline items (unknown items, missing sensors, the dead sensor, and questions such as "What did Mickey Mantle do?"), at a median of 0.31 ms per answer on a CPU. On multi-turn held-out scripts it held the thread in 55 of 58; the three failures were elliptical follow-ups ("And how many have we got?"), each declined rather than answered wrongly. The held-out phrasings were written by the same authors in the same session, so this is a weak held-out; the real test is phrasing written by others.

The same runtime with and without the world. A schema-only, zero-row world store that passes the system's build validator lets the full runtime run with every encyclopedic store detached. We ran the final admitted runtime both ways on the same 600 conversation turns and 600 human-labelled openers (Table 3). The prediction that conversation would survive almost intact was *refuted* for open-domain conversation and *confirmed* for the relational layer. Every move that asserts only about the self or the relation survived unchanged or rose (role 12 → 12, customer 58 → 58, vocative 28 → 38 fitting replies; remark, compliment, care and the social moves unchanged). Every move that cites a fact fell: disclosures 61 → 20 and chain answers 54 → 16 fitting replies, with statements, advice and definitions gone. DailyDialog and OpenAssistant turns are mostly *about* the world, so general conversation quality depends on a store; what separates cleanly is the relational self, which is the part a device needs to carry everywhere, while grounding comes from whichever store the device holds. Without the world stores the 90th-percentile latency fell from 434 ms to

42 ms. An audit of every file the detached runtime opened found a core of roughly 135–145 MB on disk (a 77 MB WordNet store, 38 MB of numerical and language libraries, and about 5 MB of system code), besides two append-only experience logs that a new device would start empty.

Final runtime (w500)	with world stores	detached
Conversation judge, good of 600 (canned excl.)	194	65
Relational fit, 600 openers	338	289
Relational fit, directive openers (of 203)	74	40
Latency p50 / p90 (ms)	26.6 / 434	22.1 / 42
Peak resident memory (MB)	534	442
Encyclopedic stores on disk	>4 GB	none

Table 3: Conversation separates from world knowledge at the relational layer, not at the level of open-domain conversation quality. Resident memory differs little because the stores are read through the operating system's page cache rather than loaded; the difference that matters for a device is disk and tail latency.

9 Discussion

Why a self term, and why it can be honest here. The relational failures could not be fixed by more knowledge, because their second pair part is a fact about the relation between speakers, not about the world. A self term supplies the missing factor. In the underlying algebra, a role is a vector bound onto the self and recoverable by unbinding, so identity survives role play by construction. The self can be honest because every self-report is a measurement (turn counts, passage-binding counts, sensor readings), which satisfies the sincerity condition for expressives without claiming human feelings.

Why the gate matters more than the gain. Every improvement in this paper was checked by something that could say no, and it said no twice: once for the false-assertion rise and once for the regression list. The instruments were checked against human labels and mutants before they judged a system, and they failed three times before they were trusted. A self-modifying system is only as trustworthy as its ability to catch itself; this paper is largely a record of that ability working.

Separation and deployment. Because the relational self lives in structure and knowledge lives in stores, a device can carry the relational core and its own site while omitting the world; what it can discuss beyond itself is then exactly what its store holds, which is the intended behaviour rather than a loss. Specialisation is additive: a specialist is the shared core and self, plus a store and a sealed panel, admitted by the same gate. This is the basis for releasing many specialists that share one identity rather than many independently trained models.

10 Limitations

The conversation judge remains lenient about relevance; human ratings are pending. Relational fit is scored against one annotated corpus of scripted English dialogue, and fit is a necessary rather than sufficient condition for a good reply. About a third of opening turns are still declined. The separation law is enforced structurally but not yet tested formally. The site results use a synthetic

site, and its held-out phrasings share authors with its development phrasings. Emotion reading from a single line is weak, and we make no claim for it. All measurements are English.

11 Conclusion

A weightless, evidence-bound system can hold a far larger share of ordinary conversation when it is given a self term: a frozen identity, a conversational memory, relational moves conditioned on the turn, and self-reports that are measurements. On sealed held-out turns, good replies rose from 6.3% to 30.0% with no rise in false assertions. The same architecture, with its world stores detached, runs a site conversation in fractions of a millisecond and declines what it cannot sense. The most important result may be procedural: every step was admitted by a gate that twice refused it, and measured by instruments that were themselves caught and corrected. That is the property the larger programme depends on.

References

[1] J. L. Austin. *How to Do Things with Words*. Oxford University Press, 1962.

[2] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

[3] J. R. Searle. A classification of illocutionary acts. *Language in Society*, 5(1):1–23, 1976.

[4] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8(4):289–327, 1973.

[5] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr and T. Johnstone (eds.), *Appraisal Processes in Emotion*. Oxford University Press, 2001.

[6] J. Weizenbaum. ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[7] Y. Li, H. Su, X. Shen, W. Li, Z. Cao and S. Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*, 2017.

[8] A. Köpf et al. OpenAssistant Conversations — democratizing large language model alignment. In *NeurIPS Datasets and Benchmarks*, 2023.

[9] E. Perez et al. Discovering language model behaviors with model-written evaluations. arXiv:2212.09251, 2022.

[10] M. Sharma et al. Towards understanding sycophancy in language models. arXiv:2310.13548, 2023.