Aleten Labs

RESEARCH PAPER 2 OF A SERIES
WP-2026-02 · V1.1
AUGUST 2026 · CC BY 4.0

Honest before capable: typed failure and repair-first selection in dialogue architecture

Why every stage of a conversational pipeline should return a result or a typed flag, why flags should outrank all content at response selection, and why the repair path should be built before the response path.

B. Harvell · Aleten Labs · research@aleten.com

4 + 1	Branch 1	0	Fenced	≤ 25%	$p \cdot c_w > c_t$
FLAG FAMILIES, PLUS INGEST PLANNED	REPAIR, OF SIX, STRICTLY FIRST	CONFIDENCE SCORES ANYWHERE	STRUCTURAL INDICATION AWAITS LICENSED RE-DERIVATION	HELD-OUT FLAG RATE, THE CONVERSATIONAL BAR	THE DOMINANCE CONDITION
specified	specified	specified	decided	decided	computed

ABSTRACT

Dialogue systems fail constantly: unknown words, fragments, unresolved reference, unrecognised intent. The engineering question is not how to eliminate these failures, which is impossible, but what the architecture does with them. We argue for a discipline with three commitments. First, every pipeline stage returns either a result or a *typed flag* drawn from a disjoint sum, carrying the evidence of what failed; no unlabelled partial result crosses a stage boundary, and no confidence score exists anywhere, because on a deterministic system a confidence score is a fabricated number. Second, raised flags outrank every content branch at response selection, so an unresolved comprehension problem becomes a clarification question rather than a fluent wrong answer; we give the decision-theoretic condition under which this dominance is optimal and show it holds for any deployment where wrong answers have consequences. Third, the repair path is built before the response path, an ordering we defend on incentive grounds: a system that is honest before it is capable can be made capable, while the reverse order rarely recovers. We distinguish detection from resolution across the pipeline of the companion specification (WP-2026-01), describe the repair surface as a replacement for the answer rather than a decoration of it, and motivate the programme with a register-proxy indication, its numbers withheld under the corpus policy pending re-derivation on a licensed spoken corpus, that the flag rate is structural rather than lexical. The claim is architectural, not stylistic: honesty here is a property of a selection function, provable and testable, not a tone of voice.

STATUS OF THIS DOCUMENT

Companion to WP-2026-01 (v1.5), which specifies the engine this paper argues design positions for. Figures carry a basis: computed, measured (cited to the engine report for build df353a9add9aa347), or decided. All figures from the unlicensed register-proxy corpus are withheld pending re-derivation on a licensed spoken corpus.

01 INTRODUCTION

failure is the common case

A conversational turn is an adversarial event. Speakers produce fragments, restarts, multi-sentence runs, words the system has never seen, referring expressions with no antecedent, and questions whose force is nothing like their form. Any pipeline that comprehends in stages will, at some stage, on a large fraction of real turns, fail to produce a full analysis; §6 reports a register-proxy indication that structural conditions implicate a substantial majority of turns against the companion engine's specified parser, with the numbers withheld pending licensed re-derivation. The design question every dialogue architecture answers, explicitly or by accident, is: **what happens to the turn when a stage cannot do its job?**

Statistical systems answer by construction: the show goes on. A neural conversational model has no stage that can fail, because every input yields a distribution over outputs; comprehension failure is not representable, so it is not reported, and the model produces its most plausible continuation of a conversation it did not understand (Bender and Koller, 2020; Vaswani et al., 2017). Fluency and comprehension are decoupled exactly where their coupling matters most. Classical pipelines often answer little better, patching failures locally with heuristic defaults so that a wrong analysis, rather than no analysis, propagates downstream unlabelled.

This paper defends a third answer, implemented in the Ackren engine (Aleten Labs, 2026), in which failure is a first-class, typed value with strict priority over success. The contributions are positions with proofs and measurements attached rather than components: the flag algebra and its two invariants (§2); the dominance argument for repair, as decision theory rather than politeness (§3); the build-order argument, that the repair path precedes the response path (§4); the distinction between detection and resolution, with a stage-by-stage accounting of which failures are which (§5); and the measured motivation, which relocated the coverage frontier from vocabulary to structure and set the numerical bar for the word "conversational" (§6). Throughout, the companion specification WP-2026-01 supplies the formal machinery; this paper supplies the arguments for having built it that way.

RELATED POSITIONS

The conversation-analysis tradition established that repair is not an exception path in human dialogue but one of its central organisations, with a documented preference structure over who initiates and who executes it (Schegloff, Jefferson and Sacks, 1977); clarification requests are frequent, conventionalised, and typed by what they target (Purver, 2004). Grounding theory models dialogue as the continuous management of mutual understanding, in which evidence of non-understanding obliges collaborative correction (Clark and Schaefer, 1989). Information-state approaches give repair a structural home: a clarification is a question pushed onto the discussion stack like any other, and an obligation to address the trouble source (Traum and Allen, 1994; Traum and Larsson, 2003; Ginzburg, 2012). What the present architecture adds to this lineage is severity and proof: repair is not one dialogue move among many but the strictly dominant branch of a provably total selection function, and every repair the system produces carries in its derivation record the typed flag that forced it. The nearest neural practice, abstention and selective answering under uncertainty estimates, shares the goal and lacks the mechanism: a calibrated probability that the system is wrong is still a guess about a guess, where a typed flag is a report of a determinate event that did or did not occur.

02 THE FLAG ALGEBRA

failure as a typed value

Let the pipeline stages be functions s_k with the convention that each returns a value in ($result_k$ + F), where

$$F = F_{oov} + F_{parse} + F_{ref} + F_{act} \quad (+ F_{input}, \text{ planned, for byte-level ingest})$$

a **disjoint sum**: a flag knows which family it belongs to, and families never overlap. Two invariants govern the algebra, and everything else in this paper depends on them.

Invariant 1 (no unlabelled partial results). A stage either completes its analysis or emits a flag; there is no third path on which a partial or guessed analysis continues downstream without a label. The parser that reaches its depth bound does not return the largest tree it happened to build as if it were the analysis; it returns F_{parse} carrying that partial structure as evidence. The distinction is between a value and a value-with-a-warrant, and downstream stages are typed to refuse the former where the latter is required.

Invariant 2 (flags carry their evidence). A flag is not a boolean. F_{oov} carries the surface form and any morphological analysis attempted; F_{parse} carries the depth reached and the spanning constituents found; F_{ref} carries the referring expression and the candidate list that failed to resolve it, or the multiple candidates that ambiguously licensed it; F_{act} carries the layers tried. The evidence is what makes the repair *constructible*: the clarification question of §3 names what failed because the flag delivered the material to name it with. A system whose failure signal is a bit can only say "I did not understand"; a system whose failure signal is a value can say which word, which reading, which referent.

2.1 · WHY THERE ARE NO CONFIDENCE SCORES

Nothing in the pipeline emits a probability of being right. This is a considered exclusion, not an omission. The engine's stages are deterministic: on given input against a given artefact, an analysis either exists under the rules or does not, and a number between 0 and 1 attached to that fact would measure nothing. Confidence scores in deployed systems are routinely uncalibrated theatre, and worse, they convert a typed, evidenced event into a scalar that downstream logic will inevitably threshold, reintroducing by the back door exactly the silent-guess behaviour the algebra exists to prevent. The output type of every stage is **resolved or flagged, never probable**. Where genuine ambiguity survives the rules, the flag says so and carries both readings; picking one by score would be a silent guess with a decimal point.

2.2 · PROPAGATION

Flags accumulate: the set f ⊆ F raised across stages 1 to 4 travels with the turn into selection. The propagation rule is total dominance, stated in §3, but one property matters before selection is even reached: flags are **not consumed en route**. A later stage may add information (the parser may confirm that the OOV token occupied a subject position) but may not discharge a flag, because discharge is a response-level decision. The one systematic exception is benign: lexical ambiguity from stage 1 is resolved *by* stage 2 when exactly one reading unifies, which is not a discharge of failure but the normal division of labour, and it is invisible in F because ambiguity-pending-parse was never flagged.

> ⊢ **KEEP** The algebra in one sentence: failure is data, with a type, carrying evidence, immune to silent discharge, and racing nothing, because it always wins.

03 WHY REPAIR DOMINATES

decision theory, not politeness

The companion specification defines selection as a strict priority stack: repair if any flag is raised, else obligations, else the question under discussion, else domain initiative, else topic continuation, else an explicit fallback, with totality and uniqueness proved (Aleten Labs, 2026). This section defends the top of that stack.

Let p be the probability that a raised flag marks a genuine comprehension gap, c_w the cost of acting on a wrong analysis, and c_t the cost of one clarification turn. Answering on a flagged analysis has expected cost $p \cdot c_w$; repairing costs c_t with certainty. Repair is the better policy exactly when

> $$p \cdot c_w > c_t$$

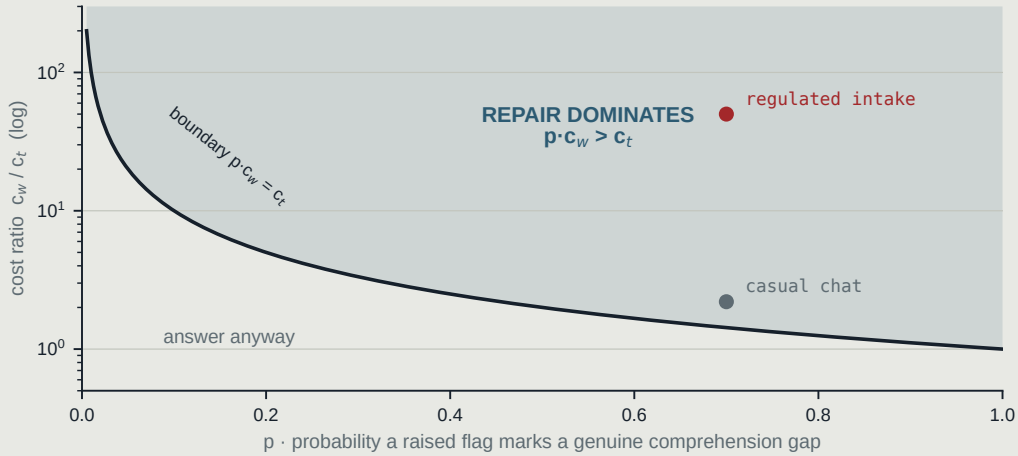

Figure 1 · The dominance region. Above the boundary $p \cdot c_w = c_t$, repairing is optimal. In casual conversation (cost ratio near 2) repair pays only when flags are highly reliable; in regulated intake (ratio 50 and far beyond) repair dominates for any plausible p. Ackren's flags fire on determinate structural failure, not on noise, so p is high by construction, and the engine's deployments sit in the upper region by definition of the market. · `Basis: computed.`

Three observations complete the argument. First, p is not a free parameter to be estimated charitably: because flags fire on structural events (no parse exists, no referent agrees, no act pattern matches), a raised flag is close to a proof of a gap, and false flags are grammar-authoring defects with a backlog entry, not ambient noise. Second, c_w is not symmetric with c_t even in casual use, because a wrong answer costs the turn it wasted *plus* the detection and unwinding turns that follow, plus trust, and trust does not regenerate at one unit per turn. Third, the clarification the system asks is not a generic "please rephrase": by Invariant 2 it names the trouble source, which conversation-analytic work identifies as the cheapest repair format for the human to process (Schegloff, Jefferson and Sacks, 1977; Purver, 2004), so c_t is at the low end of its range precisely because the flags are typed.

3.1 · THE PERCEIVED-COMPETENCE COROLLARY

A system that reliably signals its own limits reads as more competent than one that guesses fluently and is sometimes wrong. This is the folk version of the calculation above, and it has a mechanism: each fluent wrong answer is evidence to the user that fluency carries no information about correctness, after which *every* answer, right or wrong, is discounted. The flagging system never spends that capital; its confident answers stay creditable because its non-answers are visible. The repair path is therefore not a cost centre with a safety justification, it is where the product's credibility is manufactured.

3.2 · THE REPAIR SURFACE REPLACES THE ANSWER

A consequence taken seriously in the interface: when a flag forces repair, there is no answer to decorate. The engine does not render a best-guess reply with an amber banner above it, because the banner pattern asserts the guess and disclaims it in the same breath, and users act on the guess. The repair *is* the turn's entire output: a clarification question, an admission of not following, or a question about the unknown term, constructed from the flag's evidence and logged with the flag that forced it. Where the engine is permitted to assume (a licensed inheritance default, marked in the surface form and entered in the assumption ledger), that is a different, gentler mechanism with its own revision policy, and it is never available for the structural failures F was built for.

04 THE BUILD-ORDER ARGUMENT

repair before response

The engine's build sequence places the repair path before the normal response path: selection's flag branch is implemented and tested while the content branches below it are stubs. We defend this as an engineering discipline with an incentive-theoretic core.

The incentive argument. A team that builds capability first acquires, from its first demo onward, a constituency for the answer: metrics, stakeholders and habits that reward the system saying *something*. Honesty retrofitted into such a system is a regression on every dashboard, because turns that previously produced answers now produce questions, and the pressure to threshold, soften or bypass the new flags is continuous and well-meaning. Built in the other order, the system's baseline behaviour is the honest question, capability lands as pure improvement, and no dashboard ever frames a flag as a loss. The ordering is a commitment device: it makes the cheap path and the honest path the same path.

The debugging argument. Failure paths exercised only by failure are the least-tested code in any system. Building repair first inverts the exposure: for the whole early life of the engine, *every* turn traverses the flag machinery, which is therefore the most-exercised code by the time capability arrives. The specification's ordering of bridging late (step 9 of 10) is the same logic applied to knowledge: until the parser is solid, a bad bridge and a bad parse are indistinguishable in the logs, so the honest layer is stabilised before the clever one is attempted.

The honest-deferral rule. Repair extends beyond the turn. When the system lacks a knowledge cell and defers ("I will find out"), the deferral is made mechanically true: the gap is written to the annotation backlog, verified by a human, and enters the shared knowledge base in a later build, and the reply says exactly that. Phrasings that assert a nonexistent process ("give me a second while I research that") are banned as claims about the mechanism that are false. When the user supplies the missing fact, it is acknowledged, adopted session-scoped with status `USER_ASSERTED`, marked in the assumption ledger, and queued for verification with provenance back to the conversation that taught it; nothing enters the shared knowledge base without a human judgement. The machine may lean; only people may write.

> ⊢ **KEEP** The slogan, unpacked: honest before capable is not a virtue claim, it is a build order, a test-exposure strategy, and a commitment device against the one pressure every conversational product faces, which is to keep the answer flowing.

05 DETECTION IS NOT RESOLUTION

a stage-by-stage accounting

The architecture's honesty rests on a distinction that summaries of it must preserve: several stages can **detect** a condition they cannot **resolve**, and the flag algebra is what lets detection be shipped as a feature with defined behaviour rather than deferred until resolution exists. The accounting:

STAGE	DETECTS AND RESOLVES	DETECTS, NEVER RESOLVES	REPAIR RENDERED AS
`1 · Analyse`	known words, licensed affixed forms	out-of-vocabulary tokens; the lexicon cannot contain English	a question about the term, quoting it via the typed Quoted channel
`2 · Parse`	ambiguity that unification eliminates; attachment by fixed preference	fragments and multi-sentence turns beyond the grammar (until the structural milestone); genuine structural ambiguity, flagged with both readings	an admission of not following, carrying the largest constituents found
`3 · Refer`	pronouns and definites with agreeing antecedents; lexically licensed bridges	novel bridging (coverage < 1 strictly); ambiguous licence	a clarification naming the candidates
`4 · Act`	literal and conventionalised force	non-conventionalised implicature: detected as a literal answer that would be uninformative, unresolvable without open-domain goal inference	answering the literal act while flagging the oddity in the log
`Knowledge`	authored cells; licensed inheritance defaults, marked and ledgered	gaps relative to truth: the engine detects gaps relative to what has been authored, never relative to what is true	the honest deferral of §4, or an ask-only question where sibling structure makes ignorance conspicuous

The last row's ask-only constraint deserves its sentence: where the knowledge structure makes a gap conspicuous (peer concepts carry a slot this one lacks), the engine may **ask** about the gap, and may never **assert** a filler for it, because sibling patterns license curiosity and do not license inference; asserting from them would rebuild by hand the co-occurrence failure mode the knowledge layer explicitly rejects `(Aleten Labs, 2026)`. Detection without resolution, surfaced as a question, is the same repair discipline applied to knowing rather than parsing.

Two failure modes remain outside the accounting entirely, disclosed in WP-2026-01's failure table: index false accepts (probabilistic, bought down with fingerprint width) and long-range incoherence (undetectable within the design). A summary of this architecture that says "it knows what it doesn't know" is therefore wrong twice over, and the accurate sentence is the one the programme uses: **it detects gaps relative to what has been authored, never relative to what is true.**

06 THE PROVISIONAL MOTIVATION

the flag rate is structural; numbers withheld

The empirical motivation for this paper's positions comes from a register-proxy analysis: the specified pipeline's flag conditions run over a 2006 web-scraped quotation corpus distributed as an NLTK teaching sample, whose licence is unclear and whose register is edited quotation rather than spontaneous speech. Under the programme's corpus policy an unclear licence is a denial, so none of that analysis's numbers appear here; they will be published when re-derived on a licensed spoken corpus, a scheduled milestone, whichever way they land. A correction is recorded for the same reason the architecture records its own: an earlier draft described this analysis as a field study of collected turns, and no collection event occurred. What the proxy indicated, stated qualitatively: multi-sentence turns and verbless fragments implicate a substantial majority of turns, while solving out-of-vocabulary entirely moves the flag rate by only a few points. Three consequences follow, and none depends on the withheld numbers.

First, the repair path is the primary user experience for the engine's whole early life, which retroactively strengthens §4: the path users actually traverse is the one built and polished first. **Second, "fully conversational" is retired as a claim.** Open-domain vocabulary is asymptotic and open-domain structure is a research programme, so the phrase promises a limit the curves say is unreachable; the replacement claim is behavioural and checkable: *sound, incomplete, and honest about which of the two it just hit, with a derivation log a person can re-derive by hand.* **Third, the word "conversational" acquires a number.** Two published bars, each to be measured on held-out turns from a licensed corpus, replace adjectives: CONVERSATIONAL at an in-scope flag rate ≤ 25%, SCOPED-COMPLETE at ≤ 15%. A bar the engine has not passed is stated as unpassed. This is the repair discipline applied to the marketing department, and it is the only version of the claim a design partner can hold the programme to.

6.1 · WHAT WOULD FALSIFY THE POSITION

The dominance argument fails empirically if measured p is low, that is, if a large share of raised flags mark turns the system would have answered correctly; the falsification test is a blind study scoring repair appropriateness against annotator judgement of whether a gap existed, and it is in the acceptance milestone. The build-order argument fails if the repair-first engine proves unable to grow capability at a competitive rate, which the milestone gates make measurable. The authors' position is that both tests should be run and published whichever way they land; an honesty architecture that would not publish its own falsification results would be a contradiction in terms.

07 LIMITS AND CONCLUSION

what the discipline does not buy

Limits. The flag algebra types the failures the pipeline can witness; it is silent on the two failures WP-2026-01 marks unsound (index false accepts, long-range incoherence) and on wrongness of the knowledge base itself, where a deterministic engine is a machine for giving the same wrong answer to every user, with a receipt. Repair dominance is optimal under the cost model of §3 and that model's costs are deployment-specific; in genuinely low-stakes settings the boundary is real and a deployment may legitimately tune the assumption mechanism, though never the structural flags, toward leaning further. And the honesty of the whole construction is conditional on the verification programme that audits what "authored" means; typed flags over an unverified knowledge base would be a precise account of distance from an unchecked map.

Conclusion. We have argued that a dialogue system's treatment of its own failures is an architectural decision with a provable shape: failure as typed, evidence-carrying values; repair as the strictly dominant branch of a total selection function, optimal wherever wrong answers cost more than questions; the repair path built first, as a commitment device; detection shipped honestly where resolution is impossible; and the whole discipline calibrated by a measurement showing that repair, not fluent success, is the common case of real conversation. None of this required a new algorithm. It required deciding, before capability existed to tempt anyone otherwise, that the system would never be fluent about what it had not understood.

A REFERENCES

BU Harvard, alphabetical

Aleten Labs (2026) *Ackren: a formal specification of a deterministic conversational engine with per-response derivability*. Working paper WP-2026-01. Aleten Labs.

Allen, J.F. and Perrault, C.R. (1980) 'Analyzing intention in utterances', *Artificial Intelligence*, 15(3), pp. 143–178.

Austin, J.L. (1962) *How to do things with words*. Oxford: Oxford University Press.

Bender, E.M. and Koller, A. (2020) 'Climbing towards NLU: on meaning, form, and understanding in the age of data', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 5185–5198.

Clark, H.H. (1975) 'Bridging', in *Theoretical Issues in Natural Language Processing (TINLAP-1)*, pp. 169–174.

Clark, H.H. and Schaefer, E.F. (1989) 'Contributing to discourse', *Cognitive Science*, 13(2), pp. 259–294.

Fleiss, J.L. (1971) 'Measuring nominal scale agreement among many raters', *Psychological Bulletin*, 76(5), pp. 378–382.

Ginzburg, J. (2012) *The interactive stance: meaning for conversation*. Oxford: Oxford University Press.

Grice, H.P. (1975) 'Logic and conversation', in Cole, P. and Morgan, J.L. (eds) *Syntax and semantics 3: speech acts*. New York: Academic Press, pp. 41–58.

Grosz, B.J., Joshi, A.K. and Weinstein, S. (1995) 'Centering: a framework for modeling the local coherence of discourse', *Computational Linguistics*, 21(2), pp. 203–225.

Purver, M. (2004) *The theory and use of clarification requests in dialogue*. PhD thesis. King's College London.

Pustejovsky, J. (1995) *The generative lexicon*. Cambridge, MA: MIT Press.

Roberts, C. (2012) 'Information structure in discourse: towards an integrated formal theory of pragmatics', *Semantics and Pragmatics*, 5(6), pp. 1–69.

Schegloff, E.A., Jefferson, G. and Sacks, H. (1977) 'The preference for self-correction in the organization of repair in conversation', *Language*, 53(2), pp. 361–382.

Searle, J.R. (1975) 'Indirect speech acts', in Cole, P. and Morgan, J.L. (eds) *Syntax and semantics 3: speech acts*. New York: Academic Press, pp. 59–82.

Traum, D.R. and Allen, J.F. (1994) 'Discourse obligations in dialogue processing', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 1–8.

Traum, D.R. and Larsson, S. (2003) 'The information state approach to dialogue management', in van Kuppevelt, J. and Smith, R.W. (eds) *Current and new directions in discourse and dialogue*. Dordrecht: Kluwer, pp. 325–353.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', in *Advances in Neural Information Processing Systems 30*.

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