Aleten Labs

RESEARCH PAPER 1 OF A SERIES
WP-2026-01 · V1.5
AUGUST 2026 · CC BY 4.0

Ackren: a deterministic conversational engine with per-response derivability

Formal architecture, quantitative model, and computed results for a dialogue system in which every output is accompanied by a machine-checkable derivation, and any output can be re-derived by hand.

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

8	80.8 MB	43.8 KB	1 read	D = 0.756	0.12 ms
TYPED STAGES, RESULT OR FLAG	FULL COMPILED ARTEFACT	SRAM, PLUS MAPPED INDEX	PER TOKEN, CO-LOCATED	KNESER–NEY DISCOUNT	REFUSAL TURN, END TO END, DEV BUILD
specified	computed	computed	computed	computed	measured

ABSTRACT

Statistical conversational systems cannot explain their outputs: interrogating a neural language model yields a plausible account, not a mechanism. We specify Ackren, a conversational engine that inverts the problem by making the explanation identical to the execution. The engine is a composition of eight deterministic stages, each a typed function returning either a result or a typed flag; no unlabelled partial result crosses a stage boundary, and raised flags dominate response selection, so every unresolved comprehension problem surfaces as a clarification question rather than a wrong answer. Realisation intersects a frame automaton compiled from the selected intent with a weighted n-gram automaton, so non-adherent output is not in the language rather than merely improbable, and a round-trip verification stage re-parses each output and proves it recovers the planned intent, logging the proof against a content-addressed artefact hash. We give the formal model, including a totality and uniqueness proof for the selection function and a strict coverage bound for bridging inference, and we derive the complete resource envelope from a designed 1.8×10^7 token corpus: an 80.8 MB artefact, 43.8 KB of working memory plus a memory-mapped index, and one block read per generated token after backoff co-location. A third of the specification is failure accounting: eleven enumerated failure modes, of which nine are detected and typed, and two are unsound and stated as such. The system is designed to be executed by hand from printed tables, so a regulator can independently re-derive any output it has ever produced. Every figure carries a declared basis: computed, measured, or withdrawn. This paper is a specification with computed results; it describes a design that can be checked, not a system that has been benchmarked.

STATUS OF THIS DOCUMENT

A specification is not an engine. Figures are labelled by basis: computed from the corpus model, measured (cited to a dated log or to the engine report for build df353a9add9aa347), or withdrawn. v1.5 additionally withdraws every figure derived from the unlicensed register-proxy corpus and corrects its earlier description as a field study (§6.8); v1.4 struck an earlier latency and an earlier held-out hit rate pending re-derivation. Measured claims cite the engine report for build df353a9add9aa347. Superseded claims are struck, never deleted. Language-statistical figures from that build's plumbing-register corpus are quarantined and never quoted here.

§ CONTENTS

ten sections, one appendix

01 INTRODUCTION

the mechanism claim

Ask a neural conversational system why it produced an output and the answer is another output. The model that generated the sentence is the only witness to its generation, and its testimony is a post-hoc narrative sampled from the same distribution that produced the sentence in the first place. This is not a defect of any particular model but a property of the architecture class: a transformer's computation is a sequence of continuous mixtures with no discrete record of which rule fired, because there are no rules to fire (Vaswani et al., 2017; Bender and Koller, 2020). For most applications this does not matter. For deployments governed by IEC 62304, DO-178C, ISO 26262 or EN 50128, all of which assume requirements traceability from output back to specified behaviour (IEC, 2006; RTCA, 2011; ISO, 2018), it is disqualifying, and Article 12 of the EU AI Act now imposes record-keeping obligations on high-risk systems that statistical architectures must satisfy by retrofitted approximation (European Union, 2024).

Ackren is a conversational engine designed so that the explanation of an output is the execution that produced it. The engine is deterministic end to end: a composition of eight typed stages in which every stage returns either a result or a typed flag, every response is selected by a provably unique priority function, every surface string is generated inside a formal language fixed by the selected intent, and every output is re-parsed by the system's own comprehension pipeline to prove it recovers the intent it was generated from. The record of one turn through this pipeline is fourteen fields written on the response path, keyed to a content-addressed artefact hash. The claim defended in this paper is a claim about mechanism, not performance: we do not argue that Ackren converses better than a neural system, we argue that it can prove what it did. Stated as a contribution: **a conversational architecture whose runtime is a proof object**, in which producing the reply and producing the evidence for the reply are the same computation, cheap enough to run on a microcontroller and simple enough to check by hand.

CONTRIBUTIONS

1. **A fully typed failure architecture** (§3, §4). Eight stages as typed functions; a disjoint flag sum F; the invariant that no unlabelled partial result crosses a stage boundary; and a selection function σ over information state and raised flags that is total, deterministic, and provably unique per state, with repair dominating all content branches on decision-theoretic grounds.
2. **Adherence by construction** (§4.3). Realisation as the intersection $A_{\text{frame}} \cap A_{\text{ngram}}$: the frame automaton fixes illocutionary shape and slot structure, the n-gram automaton weights lexical choice, and output violating the intent is unreachable rather than detected.
3. **Round-trip verification** (§4.4). $\nu(y, i) = [(\alpha \circ \rho \circ \pi \circ \tau)(y) = i]$, a per-response adherence proof, with its honest limit stated: soundness with respect to the grammar, not with respect to English.
4. **Bridging as a licence predicate** (§4.5). Definite reference without an antecedent resolved through Pustejovsky qualia roles (Pustejovsky, 1995), with a strict coverage bound and the argument that soundness with detected incompleteness is the strongest claim available to any system.
5. **A complete computed resource envelope** (§6). Every sizing figure derived from a designed corpus model and stated with its basis, including two inverted intuitions: the fingerprint costs ten times the hash it protects, and the small build is vocabulary-limited.
6. **Hand executability** (§9). The runtime is table lookup, integer comparison and bounded search, so a printed artefact and a pencil re-derive any output, at roughly 15 to 20 pencil operations per word.

Nothing in Ackren's component inventory is new. This section is therefore not a survey of adjacent work but an accounting of debts: each subsection names the results the engine is built from and states what the engine takes from them. The composition, we argue, is where the contribution lies, and a reader who finds a component treated unfairly here will find the primary source cited beside it.

2.1 · DIALOGUE AS INFORMATION STATE

The engine's discourse layer is the information-state approach: dialogue context as a structured state updated by rule as acts are recognised, with response selection reading the state rather than classifying the utterance (Traum and Larsson, 2003). Ackren's state is the four-component tuple $S = \langle Q, O, R, e \rangle$. The question-under-discussion stack follows the QUD tradition of Ginzburg and Roberts, in which asking pushes a question and only an entailing answer discharges it (Ginzburg, 2012; Roberts, 2012). The obligation queue follows Traum and Allen's discourse obligations: a greeting obliges a return, a request obliges acceptance or refusal, and obligations are objects distinct from questions (Traum and Allen, 1994). Clarification behaviour, the engine's primary output under failure, is grounded in Purver's study of clarification requests in human dialogue, which established both their frequency and their conventional forms (Purver, 2004).

2.2 · SPEECH ACTS AND PLAN INFERENCE

Act recognition is layered exactly as the literature developed it. Literal force from sentence type descends from Austin's performatives (Austin, 1962); conventionalised indirection ("can you pass the salt" as request, not question) is Searle's analysis, implemented as authored grammar patterns rather than inference (Searle, 1975); and the third layer is Cohen, Allen and Perrault's plan-based inference, reading the speaker's act off which precondition of a domain plan the utterance queries (Cohen and Perrault, 1979; Allen and Perrault, 1980). The engine inherits the known complexity consequence: backward chaining over a plan library is $O(b^d)$ in branching factor and depth, tractable on bounded domains and unusable on open ones, which is why layer three ships only inside domain packs (§4.2). Grice's implicature marks the boundary of what the engine detects but does not resolve (Grice, 1975).

2.3 · REFERENCE, COHERENCE AND LEXICAL KNOWLEDGE

Referent salience is centering theory operationalised: grammatical role ordering, backward-looking centres, and transition types as coherence signals (Grosz, Joshi and Weinstein, 1995). Bridging reference, the definite description with no antecedent, is Clark's problem (Clark, 1975), given empirical shape by Poesio and Vieira's corpus studies, which found that roughly half of definite descriptions in naturally occurring text are not anaphoric, making bridging a frequency phenomenon rather than a corner case (Poesio and Vieira, 1998). Ackren resolves the lexically licensed subset through the qualia structure of Pustejovsky's generative lexicon, whose four roles (constitutive, formal, telic, agentive) supply the relation inventory the licence predicate quantifies over (Pustejovsky, 1995). Situationally licensed reference beyond the lexicon ("the waiter" after "the restaurant") is script knowledge in the sense of Schank and Abelson, authored as scripts rather than inferred from co-occurrence (Schank and Abelson, 1977).

2.4 · DETERMINISTIC PARSING AND FINITE-STATE MORPHOLOGY

The grammar backbone is generative in Chomsky's original sense of an explicit rule system (Chomsky, 1957), parsed deterministically in the tradition Marcus established: commit at each choice point under a fixed policy, bound the stack, and fail rather than backtrack unboundedly (Marcus, 1980). Agreement and subcategorisation are enforced during construction by unification over feature structures (Shieber, 1986), implemented as union-find with the near-linear complexity Tarjan proved (Tarjan, 1975). Morphological analysis is Koskenniemi's two-level model, deterministic finite-state transduction between lexical and surface form (Koskenniemi, 1983), with the transducer calculus standardised by Mohri (Mohri, 1997). English-specific ordering constraints, including the conventional prenominal adjective order documented by Quirk et al. and given a behavioural basis by Scontras et al., live in the grammar rules of the language artefact, not in the engine (Quirk et al., 1985; Scontras, Degen and Goodman, 2017); this division is developed in §3.

2.5 · STATISTICAL LANGUAGE MODELLING, FROZEN AT BUILD TIME

Lexical choice inside grammatical structure is ranked by an interpolated Kneser–Ney n-gram model (Kneser and Ney, 1995), the strongest of the classical smoothing family in Chen and Goodman's systematic comparison (Chen and Goodman, 1999), with unseen-mass estimation due to Good (Good, 1953). The distributional facts the sizing model rests on, type growth and rank-frequency skew, are Heaps' and Zipf's laws respectively (Heaps, 1978; Zipf, 1949), and §6.1 shows why the first, unusually, does not apply to this corpus. The essential point for determinism is that all counts, discounts and interpolation weights are computed once at build time and frozen into the artefact: at runtime there is no sampling and no estimation, only deterministic search over stored weights, which is why the engine can be statistical in provenance and deterministic in execution.

2.6 · SUCCINCT INDEXING

The artefact is indexed by minimal perfect hashing, whose modern constructions approach the information-theoretic bound of 1.44 bits per key (Botelho, Pagh and Ziviani, 2007; Esposito, Mueller Graf and Vigna, 2020), with constant-time rank support from succinct bit-vector techniques (Jacobson, 1989). An MPHF has no membership test: an absent key hashes to some occupied slot, so every payload carries a fingerprint whose width is an explicit false-accept parameter, the same space-for-certainty trade Bloom analysed for approximate membership (Bloom, 1970). Section 6.6 shows that the fingerprint, not the hash, dominates the index budget, and treats its width as an auditability decision.

2.7 · THE NEURAL CONTRAST, STATED CAREFULLY

Conversational systems have alternated between authored and statistical extremes since ELIZA demonstrated how little machinery fluent-seeming dialogue requires (Weizenbaum, 1966). Transformer language models (Vaswani et al., 2017) sit at the statistical extreme, and two lines of critique frame the gap Ackren occupies: Bender and Koller's argument that form-only training cannot ground meaning (Bender and Koller, 2020), and the documentation critique of scale-first systems (Bender et al., 2021). The nearest neural relatives of §4.3 are constrained decoding methods, which mask a model's output distribution against a grammar at inference time (Scholak, Schucher and Bahdanau, 2021; Willard and Louf, 2023). The distinction matters: constrained decoding restricts a sampler whose state remains uninspectable, so it guarantees well-formedness of the string but not derivability of the choice. Ackren's guarantee is the stronger one because there is nothing else in the system: the grammar path is the entire computation, not a filter on it.

2.8 · THE REGULATORY FRAME

The deployment context is set by the EU AI Act's record-keeping and transparency obligations (European Union, 2024) and by the safety-critical software standards whose shared assumption is traceability from output to requirement (IEC, 2006; RTCA, 2011; ISO, 2018). The human-verification methodology of §5.3 rests on Fleiss' kappa for multi-rater agreement (Fleiss, 1971). None of these instruments requires determinism by name; all of them require properties that determinism supplies and statistical inference does not, which is the precise form of the regulatory claim this paper permits itself.

2.9 ORIENTATION

for readers arriving from outside computational linguistics

2.9.1 · MECHANISM COMPARISON, NOT A QUALITY RANKING

PROPERTY	LLM, UNCONSTRAINED	LLM + CONSTRAINED DECODING	CLASSICAL RULE PIPELINE	ACKREN
Deterministic execution, per build	no; sampled	no; sampled within the mask	usually	`yes, by construction`
Complete per-response derivation	no	no; the sampler's state is uninspectable	partial; rule traces where implemented	`yes: 18-field record on the response path`
Output structure guaranteed	no	well-formedness of the string only	template-bound	`in the language of the intent frame`
Comprehension failure signalled	no; fluent continuation regardless	no	often silent or brittle	`typed flags; repair outranks content`
Independent re-derivation, incl. by hand	no	no	rarely practical	`yes, from the printed artefact`
Knowledge provenance	parametric, unauditable	parametric, unauditable	authored	`authored, human-verified, per-item`
Open-domain coverage	**yes**	**yes**	**no**	**no: authored scope, flagged at the boundary**
Long-form coherence	**often, unverified**	**often, unverified**	**no**	**no, and undetectable: an unsound row (§7)**

Mechanism properties, not quality rankings. The two rows where Ackren is the weaker system are included deliberately; a comparison table that only contained the favourable rows would violate the discipline this paper argues for.

2.9.2 · EIGHT TERMS, TWO LINES EACH

`QUD`	questions under discussion: the stack of open questions a conversation is currently answering; asking pushes, an entailing answer pops
`qualia structure`	a lexical entry's typed relations (constitutive, formal, telic, agentive): a book has an author by its agentive role, which is what licenses "the author" after "the book"
`centering`	a theory of local coherence ranking referents by grammatical role, predicting what a pronoun refers to and how salience decays over turns
`two-level morphology`	word structure as deterministic finite-state transduction between a lexical form and a surface form: licensed affixes accepted, **goed* refused
`unification`	merging two partial descriptions into their least combined description, or failing; how agreement is enforced during parsing rather than checked after
`Kneser–Ney`	the standard smoothing method for n-gram models; here its weights are computed once at build time and frozen, so no estimation happens at runtime
`MPHF + fingerprint`	a minimal perfect hash gives every stored key a unique slot but cannot say a key is absent; the per-slot fingerprint supplies the membership check, at a chosen error width
`speech act`	what an utterance does (ask, request, assert) as opposed to its surface form; "can you pass the salt" is formally a question and functionally a request

2.9.3 · WHO THIS IS FOR

Deployments where being confidently wrong is legally or operationally unacceptable and every output must survive an audit: regulated conversational intake (financial suitability and disclosure, insurance claims, KYC, pharmacovigilance), certification-path embedded interfaces (medical, automotive, industrial), offline and air-gapped voice systems, and the standards and audit communities that need an existence proof of per-response traceability. It is not for open-domain assistance or long-form generation, and §8 says so without hedging.

Ackren processes one utterance per turn through eight stages. Each stage is a typed function returning a value in (result + F), where F = $F_{oov} + F_{parse} + F_{ref} + F_{act}$ is a disjoint sum of flag types. Three invariants hold everywhere and are load-bearing for every claim in this paper: **every lookup is verified** (a fingerprint mismatch means absent, never present), **no stage emits an unlabelled partial result**, and **flags outrank content** (any raised flag forces the repair intent at selection, regardless of queue state).

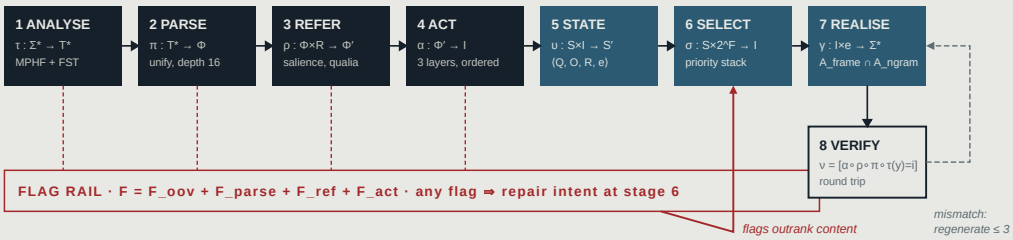

Figure 1 · The eight-stage pipeline. Comprehension stages (ink) drop typed flags onto the rail; any flag forces the repair branch of σ at stage 6, so unresolved input becomes a clarification question. Stage 8 re-enters the comprehension pipeline with the generated output and compares recovered intent against planned intent. · `Basis: specified.`

STAGE	FUNCTION	MECHANISM	TYPED FAILURES
1	Analyse · τ	Deterministic Unicode scan with authored exception list; MPHF lookup with fingerprint verify; two-level morphology on full-form miss	OOV event with surface form; ambiguous analyses passed forward unresolved
2	Parse · π	Deterministic unification parse, bounded stack (d = 16), fixed total preference order at every choice point	Stack exhaustion; no complete parse; residual structural ambiguity, flagged not picked
3	Refer · ρ	Salience-ranked referent list per centering; pronouns by agreement filter; definites by type match; bridging by qualia licence	Unresolved pronoun; unlicensed definite; ambiguous bridge (candidates attached)
4	Act · α	Ordered layers: conventionalised patterns ▷ literal force ▷ plan inference, first match wins	No act recognised; non-conventional implicature detected, not resolved
5	State · υ	QUD stack push/pop by entailment; obligation queue; salience re-rank; emotion vector as leaky integrator, read-only downstream	None: υ is total on recognised acts
6	Select · σ	Strict priority stack: repair ▷ obligation ▷ QUD ▷ initiative ▷ continuation ▷ fallback (§4.2)	None: σ is total by construction
7	Realise · γ	Intent → frame → finite acceptor; beam search over $A_{frame} \cap A_{ngram}$; closed classes templated	None reachable: non-adherent strings are not in the language
8	Verify · ν	Round trip through τ, π, ρ, α; compare recovered to planned intent; log record with artefact hash	Mismatch → regenerate (≤ 3) → closed-class fallback, logged

ENGINE AND LANGUAGE ARTEFACT: THE LAYER MODEL

The engine executes; the language artefact is executed. Everything language-specific in the table above, the lexicon, the morphological rule data, the tokeniser exception list, the grammar rules including English adjective ordering `(Quirk et al., 1985; Scontras, Degen and Goodman, 2017)`, the conventionalised act patterns, the qualia and script tables, the n-gram weights and the realisation frames, is authored data compiled into a content-addressed artefact. Beneath it sit two engine layers: a **substrate** that knows nothing linguistic (hashing, unification over an opaque feature registry, queue mechanics, intersection search, the log writer), and a thin **typology adapter** per language family, engine code that executes a family's rule shape (v1: concatenative-suffixing, configurational, article-bearing). Concept identifiers are language-free and opaque, so the knowledge layer is shared across packs. Behaviour is $E(P_L)$, and the determinism guarantee of this paper is indexed by (engine version, artefact hash). Stated honestly: within a typology family, a new language is a new artefact and zero engine change; a new family (pro-drop, non-concatenative morphology) is a new adapter, which is code. A pack itself contains no executable content, so a pack cannot alter engine semantics, which is what makes packs independently authorable, signable and certifiable.

> ⊢ **KEEP** The repair path is built before the response path: honest before capable is the order that recovers.

3.1 ONE TURN, COMPUTED IN FULL

measured, build df353a9add9aa347

The smallest complete demonstration of everything §3 specified, reproduced from the derivation log of an actual turn on the development build; nothing below is illustrative, and each step can be re-derived from the printed artefact by hand. The user asks about a word the engine does not know.

STAGE	WHAT HAPPENED, FROM THE LOG
`input`	`what is a zorblat` · 17 bytes, content-hashed into the record
`0 · ingest`	decode, whitespace map, control scan, normalisation, script scan: clean, no input flag
`1 · analyse`	known: `what (WH) · is (COP) · a (DET)`. Unknown: the fourth token. Flag raised: `{kind: ocv_oov, stage: s1, payload: {quoted_char_length: 7}}`. The unknown surface form is confined to a typed quoted span: it never enters the parse, the context, or the generation vocabulary
`2 · parse`	**not run**: analysis flagged, and parsing over unanalysed tokens would be a guess, which invariant 1 forbids
`3–5`	nothing to do on a flagged turn: no referents introduced, no act beyond the flagged path, state carried unchanged; the record says so rather than omitting them
`6 · select`	branch taken: **1, repair(f)**, because the flag set is non-empty and flags outrank content. Branches not taken, each with its failed guard: 2 obligations (queue empty) · 3 QUD (stack empty) · 4 initiative (no domain pack) · 5 continuation (referent list empty) · 6 fallback (not reached). Selected intent: `ack:intent/repair_ocv_oov`
`7 · realise`	frame `ack:cc/repair_ocv_oov`; spans: `[generated]` I don't know the word `[quoted]` zorblat `[generated]` . — every generated token checked against the enumerated output vocabulary
`8 · verify`	the reply re-enters the comprehension pipeline; recovered intent equals planned intent; `{outcome: pass, regenerations: 0}`
`reply`	**I don't know the word "zorblat".** · end to end in 0.12 ms median on the development machine

The record of this turn closes with the exportable audit object, aggregates only: one `ocv` flag, zero assumptions, one quoted span, six tokens, round trip pass. A paragraph of accounting for a seven-word reply is the intended asymmetry: the accounting is the product, and the reply that carries it costs a fraction of a millisecond. Two properties of this walkthrough generalise to every turn the engine will ever take. First, the turn is a pure function of its input, its state envelope and the content-addressed artefact, so this exact table is reproducible by re-execution, and the development build replays it byte-identically `[measured]`. Second, every row above was rendered *from* the derivation record rather than written *about* it; the walkthrough is the log, which is the sentence this paper exists to be able to say.

> ⊢ **KEEP** A reader who wants the failure case answered honestly should note what did not happen here: no guess at the unknown word's meaning, no fluent continuation past it, no confidence score, and no reply until the planned intent survived its own re-parse.

4.1 · FLAGS

A flag is an element of the disjoint sum $F = F_{oov} + F_{parse} + F_{ref} + F_{act}$. Each stage k computes a function into $(result_k + F)$, and the pipeline threads the accumulated flag set $f \subseteq F$ to selection. Disjointness is what makes the repair intent constructible: a flag carries its type and its evidence (the OOV surface form, the parse depth reached, the candidate referent list), so the clarification question names what failed. There are no confidence scores anywhere in the pipeline; on a deterministic system a confidence score would be a fabricated number, and the output type is resolved or flagged, never probable.

4.2 · SELECTION: TOTALITY, UNIQUENESS, AND WHY REPAIR DOMINATES

> $\sigma(S, f) =$ repair(f) if $f \neq \varnothing$ ▷ head(O) elif $O \neq \varnothing$ ▷ address(head Q) elif answerable ▷ initiative elif domain loaded ▷ continue($\operatorname{argmax}_{r \in R}$ sal r) elif $R \neq \varnothing$ ▷ fallback

Proposition 1 (totality and uniqueness). σ is a total function and σ(S, f) is unique for every (S, f). *Proof sketch.* The six guards are decidable (set emptiness, an entailment check against a finite QUD stack, a finite plan-library applicability test, and a finite argmax); the branches are mutually exclusive by the elif chain and exhaustive by the unconditional final branch; the one internal tie, equal salience in argmax, is broken by referent ID, a total order. ∎ This is the entire reason selection is a priority stack rather than a scored classifier: a classifier over the same inputs would need a tiebreak policy anyway and could not be explained by a single dominating condition, whereas every σ decision is explained by naming the first guard that held.

Why repair is the top branch. Let p be the probability that a raised flag marks a genuine comprehension gap, c_w the cost of answering wrongly, c_t the cost of one clarification turn. Answering on a flagged parse incurs expected cost $p \cdot c_w$; repairing incurs c_t with certainty. Repair dominates whenever $p \cdot c_w > c_t$, which holds for all but negligible p in any deployment where wrong answers have consequences, and p is empirically high because flags fire on structural failure, not on noise. The ordering is decision-theoretic, not stylistic.

4.3 · REALISATION: ADHERENCE BY CONSTRUCTION

The selected intent compiles to a syntactic frame: a constraint set fixing illocutionary shape, slot structure, agreement and register. The frame compiles to a finite acceptor A_{frame}; the frozen n-gram model is a weighted automaton A_{ngram} over the same slot-typed alphabet; generation is beam search for a low-cost path through the intersection:

> $A = A_{frame} \cap A_{ngram}$, cost = O(steps × beam × fanout), beam = 8, E[fanout] = 2.60

Every path through A satisfies the frame by construction, so output violating the planned intent is unreachable rather than improbable. The division of labour is the design rule: **grammar owns structure, n-grams own words.** This is where the engine parts company with constrained decoding over neural models (Scholak, Schucher and Bahdanau, 2021; Willard and Louf, 2023): there, the grammar masks a sampler whose internal state remains uninspectable; here, the masked search is the whole computation. The emotion vector e modifies token scores by a bounded cosine term, score = $\log P_{KN} + \mu \cdot \langle e, v(w) \rangle / (\|e\| \|v(w)\|)$, and never the language: α and σ are independent of e in their domains, which is what allows the affective layer to be developed separately.

4.4 · VERIFICATION: THE ROUND TRIP

$$\nu(y, i) = [\ (\alpha \circ \rho \circ \pi \circ \tau)(y) = i\ ]$$

The generated string y is re-analysed by the same comprehension stages that process user input, and the recovered intent is compared with the planned intent i. A match is logged as a per-response adherence proof, keyed to the artefact hash; a mismatch excludes the failing path and regenerates, at most k = 3 times, then falls back to the verified closed-class realisation for that intent, and the fallback is logged. Stated with its limit: ν proves that production and comprehension agree *within the grammar*. If π or α errs identically in both directions, the round trip passes on output a human would misread. The proof is soundness with respect to the grammar, not with respect to English, and a persistently high regeneration rate indicates an under-constrained frame, which is a build-time defect, not a runtime condition. On the development build, ν passes 20 of 20 frozen probes with zero regenerations and replays a hash-pinned ten-turn conversation byte-identically, at closed-class scope (§6.11) `[measured]`.

4.5 · BRIDGING AS A LICENCE PREDICATE, AND ITS BOUND

For an unresolved definite with head noun h, the licence predicate quantifies over the four qualia roles of the referent's type `(Pustejovsky, 1995)`:

$$\text{licence}(r, h) \iff \exists q \in \{\text{CONST, FORMAL, TELIC, AGENT}\} : h \in Q_q(\text{type } r), \quad B = \{r \in R : \text{licence}(r, h)\}$$

|B| = 1 binds and logs the licensing relation; |B| = 0 and |B| > 1 flag, with candidates attached. The mechanism is a set intersection in O(|R|), and it is fully explainable: the derivation record names which referent and which relation licensed each bind, the field no statistical architecture can populate. Let $\mathcal{B}$ be the bridging relations occurring in natural discourse and $\mathcal{Q}$ those expressible over the authored lexicon. Since bridging can rest on arbitrary world knowledge, $\mathcal{B}$ is not enumerable, and

$$\text{coverage} = |\mathcal{B} \cap \mathcal{Q}| \,/\, |\mathcal{B}| < 1 \quad \text{strictly.}$$

No quantity of authoring closes the gap. What the architecture delivers is **soundness with detected incompleteness**: every bind rests on a stated relation, and every bridge outside $\mathcal{Q}$ is flagged rather than guessed. We take this to be the strongest claim available to any referring system, and it is strictly stronger than the neural alternative, which offers neither the soundness nor the detection. Property transfer by co-occurrence is rejected as unsound: chairs have legs and tables occur near chairs, but tables have legs because the relation is authored on the table's type, not because proximity licenses inheritance; the co-occurrence intuition is admitted only as script participation `(Schank and Abelson, 1977)`, a fifth, authored role.

5.1 · THE DESIGNED CORPUS AND THE DERIVATION METHOD

All sizing results in §6 are derived from a corpus model with declared parameters: C = 60,000 conversations of W = 300 words, N = 1.8×10^7 tokens, generated by the engine's own authored lexicon and grammar (the childhood pipeline). N-gram type growth is modelled with power-law exponents in the ranges established for conversational text ($\alpha_2 = 0.72$, $\alpha_3 = 0.85$), rank-frequency skew per Zipf `(Zipf, 1949)`, and unseen mass per Good–Turing `(Good, 1953)`. Every derived figure is therefore falsifiable by construction of the artefact: compile the corpus, count, and compare. Where the development build has already yielded a measurement, the measured value supersedes the computed one and is cited to a dated development-log entry; live measurements to date comprise the mechanism results of §6.11, cited to the engine report for build df353a9add9aa347. ~~An earlier held-out trigram hit rate of 0.461~~ is struck pending re-derivation on a real build against the designed corpus; until then the trigram figures in §6.3 stand as computed. ~~An earlier end-to-end latency of 3.6 ms per 80-word reply~~ is struck pending re-derivation on the current build with the published harness; the only latency this paper quotes is the scoped refusal-turn measurement of §6.11. One prior figure, an energy comparison against hosted inference, is **withdrawn**: it rested on an unpublished batch size and is struck until it can be recomputed from published inputs.

5.2 · ARTEFACT COMPILATION

Authored sources (lexicon entries, grammar rules, qualia relations, frames, act patterns) are compiled with the frozen corpus statistics into a single packed binary: fixed-width bitfields for morphosyntactic features, shared tables for subcategorisation and qualia, 2-byte centroid references for the 12-dimensional valence vectors, minimal-perfect-hash index with per-slot fingerprints, and 512 B blocks that co-locate each context's continuations with its full backoff chain. The compile step is a measured ~60× reduction against the same data as JSON and is not an optimisation but a feasibility condition (§6.6). The artefact is content-addressed; the build identifier enters every derivation record.

5.3 · HUMAN VERIFICATION PROGRAMME

The knowledge base is verified, not trusted. Three instruments, each doing a different job: **gold items** (5% seeded known-answer items, scored sequentially under a sequential probability-ratio test with the error hypothesis measured in a 200-item pilot rather than assumed) catch inattentive work; **overlap** (triple annotation of 20% of items, Fleiss' κ `(Fleiss, 1971)`, with $\kappa < 0.60$ read as an ill-posed question rather than a failed annotator) measures whether the task is answerable; and **expert adjudication** of all disagreements and all "unsure" responses catches correlated error, which majority voting cannot: three annotators sharing a wrong intuition produce a confident wrong answer with a clean agreement score. Annotators supply judgements; a single grammar author writes rules; the derivation log names both, so a rule change traces to the evidence that motivated it. Compensation is per verified hour, never per item, so the incentive structure does not penalise honest uncertainty. Provenance is recorded per item: seeded, generated, or human-verified, with agreement score, because a knowledge base that is sound with respect to an unchecked source is not sound.

5.4 · BEHAVIOURAL TEST ITEMS AND THE REGISTER PROXY

Beyond judgement collection, the methodology uses held-out behavioural items (for bridging: "I opened the book. The author was unknown." and the check is whether the engine binds *author* through the agentive role of *book*), which test the deployed behaviour rather than a proxy and cost nothing to re-run as a regression suite. To pressure-test coverage assumptions ahead of licensed data, the programme also ran the specified pipeline's flag conditions over a **register proxy**: a 2006 web-scraped quotation corpus redistributed as an NLTK teaching sample, whose licence is unclear, whose speaker labels were assigned by anonymous submitters, and whose register is edited quotation rather than spontaneous speech. Under the programme's corpus policy an unclear licence is a denial, so no figure derived from that corpus is quoted in this paper or anywhere external; §6.8 reports the qualitative indication only, and the quantitative claim publishes when re-derived on a licensed spoken corpus, a scheduled milestone.

5.5 · BASIS DISCIPLINE

Every figure in this paper carries a basis: `computed` (arithmetic from the corpus model or from cited inputs), `measured` (run on the development build, cited to the build's engine report), or withdrawn. Corrections are first-class: a superseded figure is struck in the public development log, never deleted, and this paper's own claims are versioned under the same rule.

6.1 · THE VOCABULARY CEILING: HEAPS' LAW DOES NOT APPLY

On natural corpora, vocabulary grows as $V(N) = K \cdot N^{\beta}$ (Heaps, 1978), predicting V ≈ 158,000 at this N. That prediction is wrong here, and the reason is a design property: a self-generated corpus cannot contain a word its generator's lexicon lacks, so $V_{synthetic} = |L_{generator}|$ is a hard ceiling and V is a chosen parameter, not an empirical outcome. The consequence cuts both ways. Sizing becomes exactly predictable, but **the corpus supplies distributional statistics over existing competence, not competence**: the 60,000 conversations teach the engine how its own language is distributed, and nothing else.

6.2 · N-GRAM INVENTORY

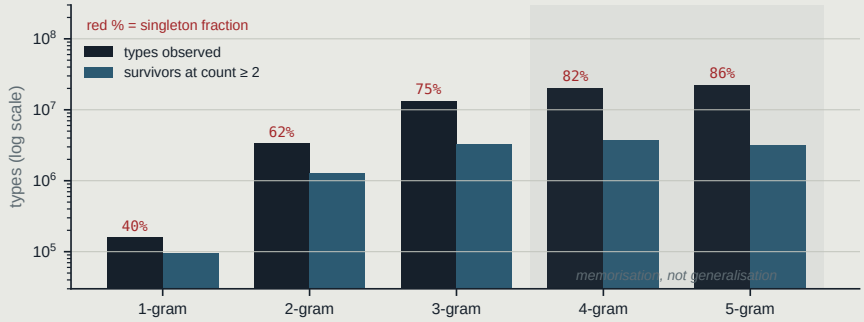

Figure 2 · Type counts by order at N = 1.8 × 10⁷, with singleton fractions (red) and count ≥ 2 survivors. Type growth saturates past n = 4 while singleton share climbs to 86%: beyond 4-grams the model stores near-unique sequences, which is memorisation, not generalisation. The build caps at order 3. · `Basis: computed.`

6.3 · COVERAGE AND BACKOFF

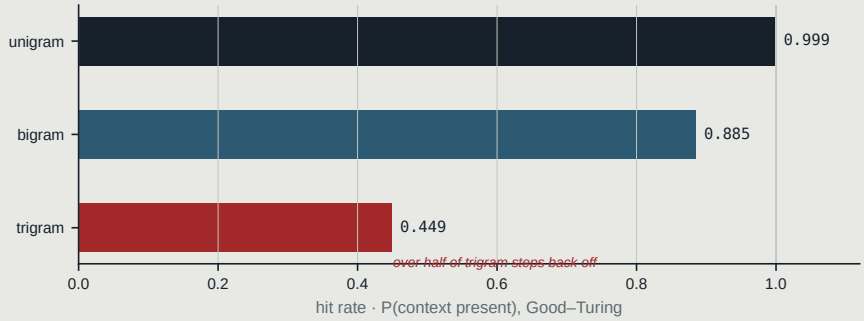

Figure 3 · Good–Turing hit rate by order. ~~An earlier held-out measurement of 0.461~~ is struck pending re-derivation; the 0.449 shown is the model's computed value. · `Basis: computed.`

Expected backoff depth E[levels] = $p_3 + 2(1-p_3)p_2 + 3(1-p_3)(1-p_2)$ = **1.89**. Co-locating each context's full backoff chain in one 512 B block collapses those 1.89 expected levels to **one physical read per generated token**, the single highest-value layout decision in the system. Pruning at count ≥ 2 removes 55.1% of trigram token mass against 11.5% of bigram: pruning largely eliminates the trigram order rather than degrading it, so a small build is honestly a bigram model with trigram exceptions.

6.4 · FANOUT AND SMOOTHING

Mean continuations per context (type average) = 3.30 × 10⁶ / 1.27 × 10⁶ = **2.60**. ~~A generation step therefore weighs one to three candidates.~~ Struck in v1.1: the development build measures token-weighted fanout at **33.2** against a type average of 1.94 in its register, so candidate-count claims must state their weighting; the load-bearing figure is physical reads, 0.91 blocks per candidates call, measured (§6.11). The Kneser–Ney absolute discount, computed from the trigram count-of-counts $n_1 = 9.91 \times 10^6$ and $n_2 \approx 1.6 \times 10^6$, is aggressive, correctly so for a corpus that is 75% trigram singletons (Kneser and Ney, 1995):

> $D = n_1 / (n_1 + 2n_2) = \mathbf{0.756}$

6.5 · INDEX INTEGRITY: THE FINGERPRINT COSTS TEN TIMES THE HASH

A minimal perfect hash over the 4.64×10^6 pruned n-gram keys costs 2.4 bits per key, 1.39 MB (Botelho, Pagh and Ziviani, 2007; Esposito, Mueller Graf and Vigna, 2020). But an MPHF has no membership test: an absent key resolves to some occupied slot with probability 1, so every payload carries a b-bit fingerprint and $P_{fa} = 2^{-b}$ per absent-key query. Expected silent false accepts over a deployment of Q queries at absent-key rate r:

> $$E[\text{false accepts}] = Q \cdot r \cdot 2^{-b}$$

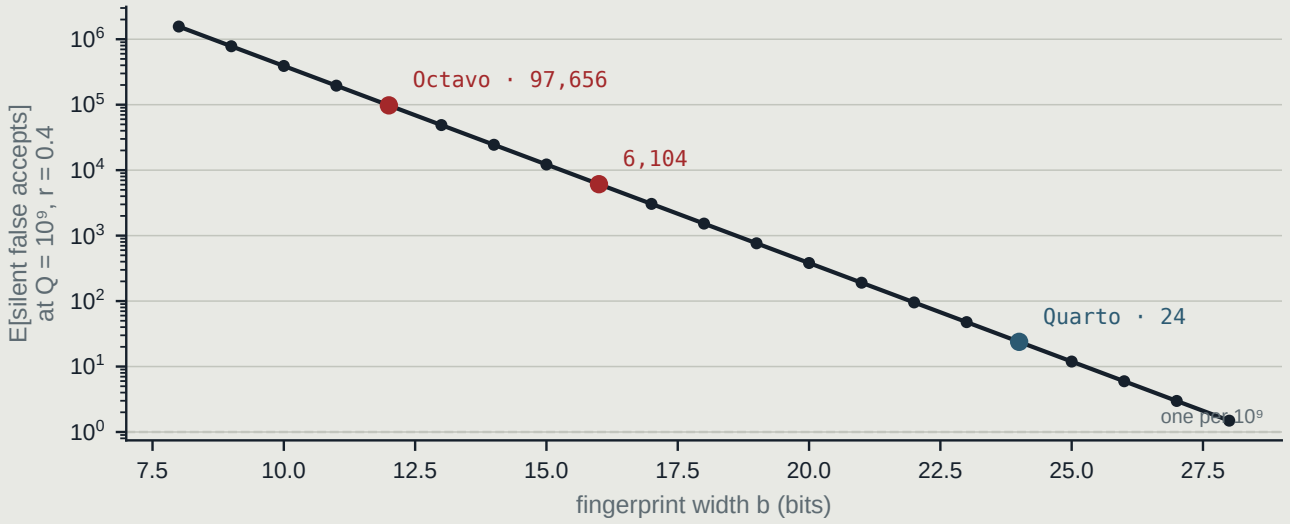

Figure 4 · Expected silent false accepts at Q = 10^9, r = 0.4, by fingerprint width. At b = 24 the index costs 13.9 MB against 1.39 MB for the hash it protects: the fingerprint costs ten times the MPHF, inverting the usual intuition that perfect hashing is nearly free. · `Basis: computed.`

Width is an auditability decision wearing a space costume: a false accept means the engine generates from an n-gram that was never in its corpus, silently, with the lookup marked verified, which is the exact failure the architecture exists to prevent. We therefore fix **b = 24 as the floor for conversational tiers** (E ≈ 24 per 10^9 queries). A 12-bit build computes to roughly five silent word substitutions per user per year under conversational load, which is why the 16 MB tier is repositioned as an embedded and demonstration tier and its weaker guarantee is stated with its name (§6.7).

6.6 · STORAGE

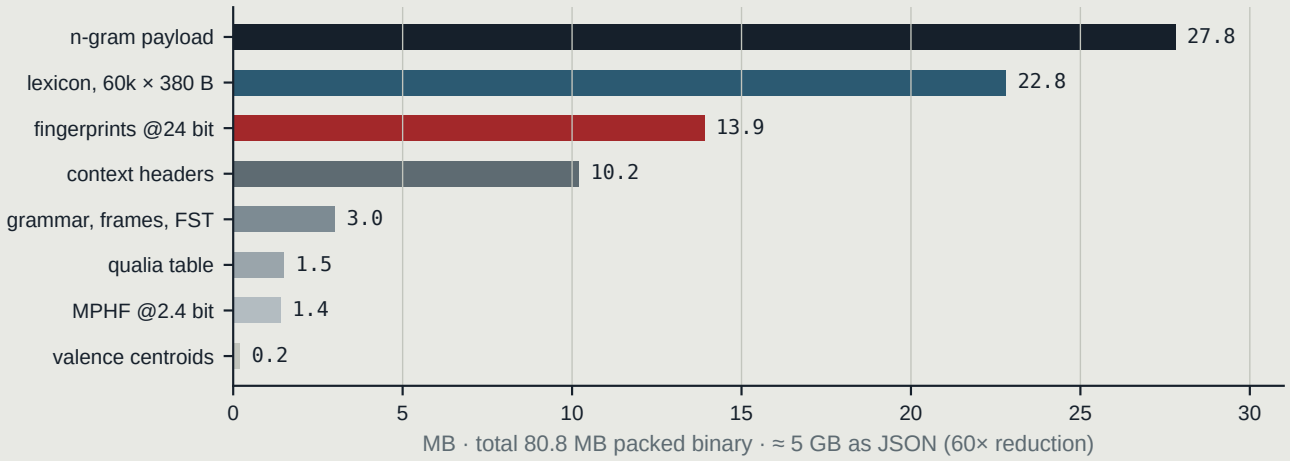

Figure 5 · The full artefact at V = 60,000, orders 1 to 3 pruned at count ≥ 2: 80.8 MB. The identical data serialised as JSON with inline float valence is ≈ 5 GB; the compile step is a 60× reduction and is the difference between embedded and impossible. · `Basis: computed.`

6.7 · WORKING MEMORY, TIERS, AND THE 16 MB FIT

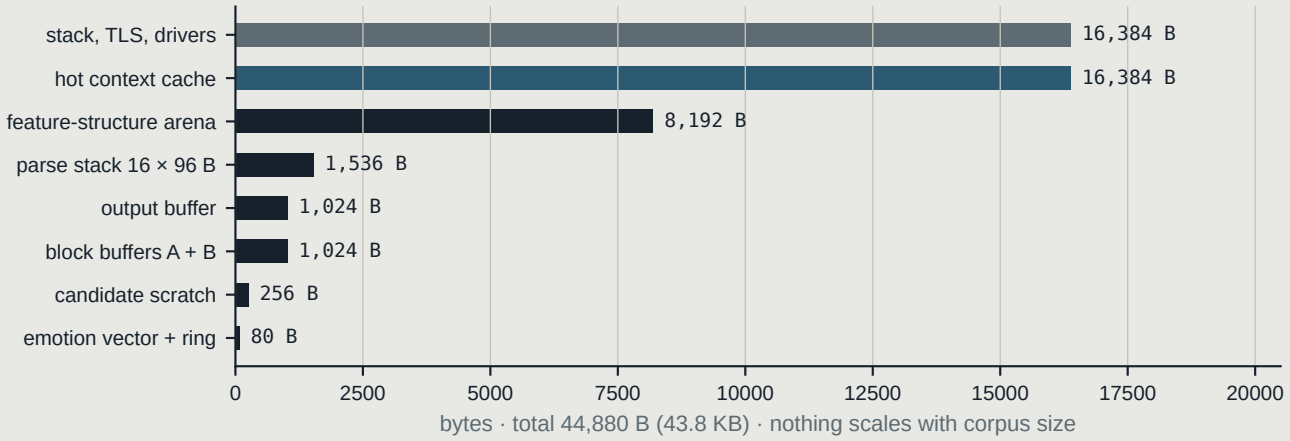

Figure 6 · The runtime SRAM budget: 44,880 bytes, with ~6 KB headroom against a 50 KB envelope. No line item scales with corpus size, which is what makes the figure robust. The caveat that must always accompany it: the MPHF and fingerprint tables do not live here; they require a memory-mapped region, so the honest claim is 50 KB of SRAM *plus a mapped index*. · `Basis: computed.`

Hardware discontinuities, not marketing, define the tiers: memory-mapped Linux (GB scale; full model, higher orders, multiple domains), PSRAM with SD storage (~200 MB; the full 80.8 MB artefact with b = 24), and 16 MB execute-in-place flash. Solving the fit inequality $6K_3 + 8K_2 + 380V + 0.3(K_2+K_3) + 1.5(K_2+K_3) \leq 16 \times 10^6$ shows the lexicon term dominating: at 380 B per entry, 8,000 entries is already 3 MB, so **the small build is vocabulary-limited, not n-gram-limited**, the opposite of the usual intuition, and the correct place to spend the next megabyte is authoring, not language-model tuning. Under the b = 24 floor of §6.5, the 16 MB tier carries a stated weaker guarantee and is offered for embedded and demonstration use, not for conversational deployments where the auditability claim is load-bearing.

6.8 · WHERE THE FLAG RATE COMES FROM: A PROVISIONAL INDICATION, NUMBERS WITHHELD

The specification's parser returns one feature structure per turn, and its lexicon budget assumed vocabulary was the binding constraint on coverage. A register-proxy analysis, run over the corpus described in §5.4, indicated otherwise: turns that are multiple sentences, and verbless fragments, jointly implicate a substantial majority of turns, while eliminating out-of-vocabulary entirely moves the flag rate by only a few points, and marginal lexicon growth buys vanishing returns. The numeric results of that analysis are deliberately not printed here: the proxy corpus carries no clear licence and is out of register for spontaneous speech, and under the programme's own corpus policy nothing measured on it may be quoted externally. The quantitative version of this claim will be published when the analysis is re-derived on a licensed spoken corpus, a scheduled milestone, whichever way it lands.

Correction, recorded here because this paper's discipline requires it: an earlier draft of this section described the analysis as a field study of collected conversational turns. No collection event occurred; the data was a pre-existing scraped corpus, and the description was wrong. What survives the correction is the design consequence, which does not depend on the withheld numbers: the build programme reordered, structural coverage before knowledge-base scale, gated on a held-out in-scope flag rate ≤ 25%; and the phrase "fully conversational" is retired in favour of the checkable claim, **sound, incomplete, and honest about which of the two it just hit**, with a derivation log a person can re-derive by hand.

6.9 · THE AUTHORING CURRICULUM AND THE STOPPING RULE

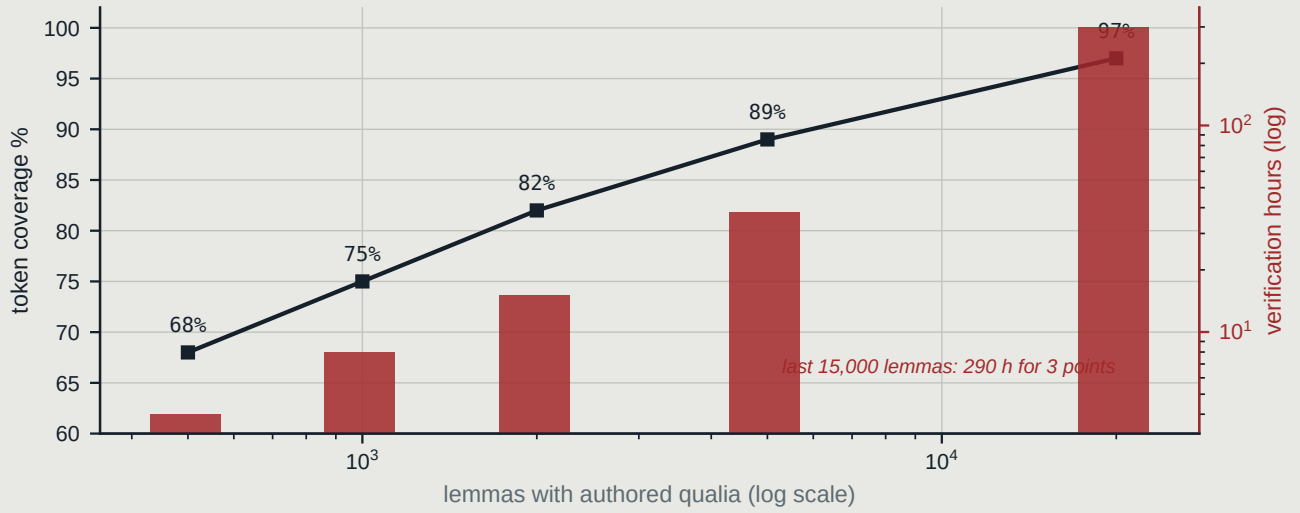

Figure 7 · Token coverage against lemmas with authored qualia (line, left axis) and verification hours (red bars, right axis, log scale). The first 1,000 lemmas cost 8 hours and cover three quarters of running text; the last 15,000 cost 290 hours for three points. · `Basis: computed.`

The curve dictates a stopping rule rather than a completion target: author in frequency order until the held-out flag rate falls below threshold, then ship, because completion is unpurchasable at any budget (§4.5's coverage bound is strict). Qualia authoring concentrates on concrete nouns, roughly 15% of a conversational lexicon; at V = 20,000 that is ~3,000 entries and ≈ 36,000 authored relations, the largest single labour item in the system, planned explicitly rather than discovered.

6.10 · LATENCY, BY STORAGE MEDIUM

MEDIUM	BLOCK READ	MS / TOKEN	TOKENS / S	NOTE
SPI SD, 512 B	2.0 ms	2.0	500	worst case considered
SDMMC 4-bit	0.6 ms	0.6	1,667	
eMMC	0.15 ms	0.15	6,667	
XIP flash / PSRAM	50–200 ns	≈ 10^{-4}	~10^{6}	embedded targets

At one read per token (§6.3), even the slowest medium exceeds conversational need (~3 tokens/s for natural pacing) by two orders of magnitude, so **latency is not the binding constraint; flash capacity is**. ~~An earlier development figure of 3.6 ms end to end for an 80-word reply~~ is struck in v1.3 pending re-derivation on the current build with the published harness; the current build measures 0.12 ms median for the refusal turn, scoped to that turn type (§6.11). The architectural point survives the stricter figure: against seconds for hosted neural systems, the gap is structural rather than tuned. Throughput per core is deliberately not quoted: it has not been benchmarked, and this paper does not print unbenchmarked speed. One unit rule accompanies any comparison: an Ackren token is one word, while neural tokenisers emit subword units, roughly 0.75 words each; we state the unit wherever a ratio is quoted.

6.11 · MEASURED MECHANISM RESULTS, DEVELOPMENT BUILD DF353A9ADD9AA347

An early development build implements all nine stages in first forms (salience-based pronoun resolution at stage 3, literal force with fragment binding at stage 4; conventionalised patterns and qualia bridging await later milestones), with closed-class realisation, a live deployed endpoint, and the full derivation-log pipeline. Its corpus is a deliberately chosen technical register (plumbing prose), so its *language-statistical* figures characterise that register and are quarantined from this paper by the build's own rule. Its *mechanism* figures, which test the machinery rather than the language, are register-safe and are reproduced here from the build's basis-tagged engine report. Two additional basis labels appear from that report's vocabulary: counted (a direct enumeration) and decided (a recorded choice).

RESULT	VALUE	SCOPE AND NOTE
Twice-built identity	`true`	two full builds in two processes yield the same content-addressed build id
Probe replay, byte-identical	`20 / 20`	frozen, hash-pinned probe set; reply and canonical log bytes both compared
Conversation replay, stateless	`10 / 10`	hash-pinned ten-turn conversation; full state carried between turns as a 1,528-byte client envelope; belief ledger exercises flip, standing negative, and attributed recall across a three-turn gap
Envelope integrity, enforced live	`true`	keyed blake2b MAC, 20-byte tag, key supplied as deployment configuration, never engine state; a deliberately forged ledger entry re-tagged with the unkeyed hash is refused by the live endpoint with a typed error
Envelope growth	`207 B · ~146 B/turn`	first turn, then mean growth over nine turns; the per-turn series plateaus because the QUD, obligation and referent caps (8/8/16) bind, so the bound is enumerated rather than discovered
Cross-build diff, first genuine entry	`2 replies changed`	diffed against the reconstructed prior build: the wh-question probe moved from a fluent non-answer to an honest deferral, the polar probe to a ledger-attributed answer; the prior was seeded by re-running the generator against the old commit, not hand-written
Round-trip verification	`20 / 20, 0 regen.`	sound with respect to the closed-class inventory at this milestone, the grammar from M4
Adversarial ingest suite	`228 / 228`	each item must raise exactly its intended flag or pass cleanly
Block reads per candidates call	`0.91`	200 calls over stride-generated contexts with a 16-block cache: the co-location design of §6.3, validated in mechanism
Reply latency, refusal turn	`0.12 ms med · 0.19 p95`	stages 0–8 plus log, 50 runs, arm64 development machine; scoped to this turn type, not a general latency claim
Deployed build integrity	`hash match`	the live endpoint reports the same build id as the local artefact; cloud evidence is display-only, since determinism proofs run locally against the pinned build
Human-verified knowledge records	**`0 / 1,143`**	**the honest state before the verification programme runs; soundness with respect to an unverified base is not yet a soundness claim**

Three of these rows earn a sentence. The stateless conversation replay demonstrates the multi-turn information-state machinery, contradiction handling included, under the strictest possible condition, the server holding nothing between turns, and the integrity of the client-carried memory is now enforced rather than assumed: the test mounts the actual forgery attack and the deployed endpoint refuses it. The cross-build diff row is the report's growth argument working as designed, and its content is worth noticing, since the first recorded behavioural change in the engine's history is two replies becoming more honest. And the zero in the final row is printed deliberately: it is the boundary between what this architecture proves about its machinery and what only the human verification programme of §5.3 can make true about its knowledge.

> ⊢ **KEEP** Every figure in this section traces to a record in the build's machine-generated report, whose generator fails on any figure without a basis tag. The paper inherits the discipline it describes.

07 FAILURE ACCOUNTING

eleven rows, two unsound

Every known failure mode, its detection condition, and whether its handling is sound. **Sound** means the engine either resolves the condition or refuses to answer; **unsound** means it can proceed while wrong. Two rows are unsound, they are the two that matter, and they are printed in proof red here rather than disclosed in an erratum. This accounting is of failure modes known to the authors; a new row appearing in a future version is the discipline working, not the discipline failing.

FAILURE MODE	DETECTION CONDITION	ROUTED TO	SOUND?
Out-of-vocabulary token	lexicon miss + fingerprint reject	repair: ask about the term	yes
MPHF false accept	**none; probabilistic, rate 2^{-b}**	**bounded by b = 24 floor, §6.5**	**no**
Parse stack exhaustion	depth counter at d = 16	repair: admit not following	yes
No complete parse	no spanning structure at root	repair, with largest constituents	yes
Residual structural ambiguity	> 1 parse after total preference order	flag with competing readings	yes
Unresolved pronoun	empty agreement-feasible set	clarification naming candidates	yes
Unlicensed bridge	|B| = 0	clarification; authoring backlog	yes
Ambiguous bridge	|B| > 1	clarification with both candidates	yes
Unrecognised act	all three α layers miss	repair	yes
Non-conventional implicature	literal answer satisfiable but uninformative	detected, flagged, unresolved	partial
Round-trip intent mismatch	ν inequality	regenerate ≤ 3, then fallback, logged	yes, w.r.t. grammar
Long-range incoherence	**none; nothing in the design detects it**	**out of scope: see §8**	**no**

Of the two unsound rows, one is purchasable: false accepts are driven arbitrarily low by spending fingerprint bits, and the b = 24 floor prices them at roughly 24 events per 10^9 queries. The other is structural: no quantity in an n-gram distribution measures whether an argument holds together across paragraphs, which is the formal reason long-form authorship is a research problem for this architecture and not a product tier. Everything else in the table is detected, typed, and routed to repair, and the enumerability of this table is itself a deployment property: enumerated, logged failure modes can be priced and underwritten, where open-ended ones cannot.

> ⊢ **KEEP** A reader who finds this table before the benefits section reaches a different conclusion about the system, and it is the conclusion we intend, because it is the accurate one. Every limit disclosed here is a limit that cannot later be discovered and presented as concealment.

08 WHAT THIS DOES NOT SOLVE

stated, not buried

1. **Novel bridging.** The licence predicate covers frequent, lexically licensed bridges. Situational and inferential bridges rest on arbitrary world knowledge, $\mathcal{B}$ is not enumerable, and coverage < 1 strictly. Detected, never resolved.
2. **Non-conventionalised implicature.** Requires modelling another mind's goals in open context; plan libraries handle bounded domains by $O(b^d)$ necessity, and the general case is open for every architecture, including neural ones, which conceal the failures this table types.
3. **Long-form coherence.** Nothing in an n-gram distribution holds an argument together across paragraphs. Undetectable within this design; authorship is a separate research problem, not a pack.
4. **Cross-build determinism.** Different lexicons and pruning yield different outputs. Determinism holds within a build, identified by hash, and we claim exactly that; the single exception is the canonical hand-execution build, required to be bit-identical across targets.
5. **Knowledge and truth.** The engine detects gaps relative to what has been authored, never relative to what is true. A deterministic system over a wrong knowledge base gives the same wrong answer to every user, with a receipt; this is why the verification programme of §5.3 is a safety control and not paperwork.
6. **Affect.** The emotional vector modulates realisation only, and affect-bearing deployment is capability-gated pending analysis of the AI Act's Article 5 prohibitions on emotion inference in workplace and education contexts `(European Union, 2024)`.

09 INDEPENDENT RE-DERIVATION

the engine on paper

The runtime consists of table lookup, integer comparison, and bounded search; there is no operation in stages 1 to 8 a person cannot perform. Print the canonical small artefact as lookup tables and a human can execute the engine: approximately 15 to 20 pencil operations per generated word, ≈ 10 minutes per word, a realistic 10 to 20 word reply in two to three hours `[computed]`. The point is not the stunt. It is that **a regulator can be handed a printout and a pencil and independently re-derive any output the engine has ever produced**, an operational definition of transparency that no statistical system can meet even in principle, and a demonstration for which the fixed, version-hashed reference artefact is a precondition. The company name commits to it: *aletheia*, unconcealment, made literal.

10 CONCLUSION

sound, incomplete, and it says so

We have specified a conversational engine whose every response carries its own derivation: a typed flag algebra with repair dominance, a provably unique selection function, realisation inside an intent-fixed formal language, and a round-trip proof logged against a content-addressed artefact. The resource envelope is computed end to end, 80.8 MB of artefact, 43.8 KB of SRAM plus a mapped index, one read per token, and the development build's measured mechanism results (§6.11), from byte-identical replay to 0.91 block reads per candidates call, landed where the model said they would. The architecture's honest ledger is a third of this paper: two unsound failure modes, one purchasable with bits, one structural; a coverage bound that no authoring closes; and a register-proxy indication, its numbers withheld under the programme's own corpus policy pending licensed re-derivation, that the coverage frontier is structural rather than lexical, which reordered our own programme. One number this paper deliberately does not contain is a conversational success rate: the programme's public bar, a held-out in-scope flag rate of at most 25%, is unrun at this writing, and it will be reported, whichever way it lands, in the empirical paper that follows the structural milestone. The claim we defend here is deliberately narrow and, we believe, currently unique: not that this system converses best, but that it is **correct where it acts, flagged where it is not, and checkable by anyone, including by hand**. Subsequent papers in this series treat the typed-flag architecture, the open derivation-record schema, the failure accounting, the vocabulary-limited small build, and the engine/language-artefact separation as standalone results.

A REFERENCES

BU Harvard, alphabetical

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.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) 'On the dangers of stochastic parrots: can language models be too big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, pp. 610–623.

Bloom, B.H. (1970) 'Space/time trade-offs in hash coding with allowable errors', *Communications of the ACM*, 13(7), pp. 422–426.

Botelho, F.C., Pagh, R. and Ziviani, N. (2007) 'Simple and space-efficient minimal perfect hash functions', in *Algorithms and Data Structures (WADS 2007)*. Berlin: Springer, pp. 139–150.

Chen, S.F. and Goodman, J. (1999) 'An empirical study of smoothing techniques for language modeling', *Computer Speech and Language*, 13(4), pp. 359–394.

Chomsky, N. (1957) *Syntactic structures*. The Hague: Mouton.

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

Cohen, P.R. and Perrault, C.R. (1979) 'Elements of a plan-based theory of speech acts', *Cognitive Science*, 3(3), pp. 177–212.

Esposito, E., Mueller Graf, T. and Vigna, S. (2020) 'RecSplit: minimal perfect hashing via recursive splitting', in *Proceedings of the Symposium on Algorithm Engineering and Experiments (ALENEX 2020)*. SIAM, pp. 175–185.

European Union (2024) *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act)*. Official Journal of the European Union, L series, 12 July.

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.

Good, I.J. (1953) 'The population frequencies of species and the estimation of population parameters', *Biometrika*, 40(3–4), pp. 237–264.

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.

Heaps, H.S. (1978) *Information retrieval: computational and theoretical aspects*. New York: Academic Press.

IEC (2006) *IEC 62304: Medical device software – software life cycle processes*. Geneva: International Electrotechnical Commission.

ISO (2018) *ISO 26262: Road vehicles – functional safety*. Geneva: International Organization for Standardization.

Jacobson, G. (1989) 'Space-efficient static trees and graphs', in *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 549–554.

Kneser, R. and Ney, H. (1995) 'Improved backing-off for m-gram language modeling', in *Proceedings of ICASSP 1995*. IEEE, pp. 181–184.

Koskenniemi, K. (1983) *Two-level morphology: a general computational model for word-form recognition and production*. Helsinki: University of Helsinki.

Larsson, S. and Traum, D.R. (2000) 'Information state and dialogue management in the TRINDI dialogue move engine toolkit', *Natural Language Engineering*, 6(3–4), pp. 323–340.

Marcus, M.P. (1980) *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.

Mohri, M. (1997) 'Finite-state transducers in language and speech processing', *Computational Linguistics*, 23(2), pp. 269–311.

Poesio, M. and Vieira, R. (1998) 'A corpus-based investigation of definite description use', *Computational Linguistics*, 24(2), pp. 183–216.

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.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A comprehensive grammar of the English language*. London: Longman.

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

RTCA (2011) *DO-178C: Software considerations in airborne systems and equipment certification*. Washington, DC: RTCA.

Schank, R.C. and Abelson, R.P. (1977) *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Scholak, T., Schucher, N. and Bahdanau, D. (2021) 'PICARD: parsing incrementally for constrained auto-regressive decoding from language models', in *Proceedings of EMNLP 2021*, pp. 9895–9901.

Scontras, G., Degen, J. and Goodman, N.D. (2017) 'Subjectivity predicts adjective ordering preferences', *Open Mind*, 1(1), pp. 53–66.

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.

Shieber, S.M. (1986) *An introduction to unification-based approaches to grammar*. Stanford, CA: CSLI Publications.

Tarjan, R.E. (1975) 'Efficiency of a good but not linear set union algorithm', *Journal of the ACM*, 22(2), pp. 215–225.

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.

Willard, B.T. and Louf, R. (2023) 'Efficient guided generation for large language models', *arXiv preprint* arXiv:2307.09702.

Zipf, G.K. (1949) *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.