Peel

Structural Hallucination Prevention for Offline AAC
Through Symbolic Fact Authorship

Perslis Research

2026-09-15

PROTOTYPE / DRAFT — unpublished.
Status: Public, moat-scrubbed research note. A patent-sensitive internal whitepaper describes the enforcement mechanisms; this document conveys the thesis, the problem, the property-level architecture, and the measured results without disclosing the protected mechanism recipes.

Abstract

Augmentative and alternative communication (AAC) systems help children and other users with complex communication needs to speak. When such a system is asked a question and answers with a fabricated fact, the failure is not a nuisance — it is a safety failure delivered in the user's own voice. Statistical language models mitigate hallucination probabilistically; they do not prevent it structurally, and probabilistic mitigation is the wrong guarantee for this population. We describe **Peel**, an offline, on-device symbolic reasoning system built on a single architectural commitment we call *the inversion*: a typed symbolic knowledge store is the **only author of facts**, and every neural model is demoted to one of two subordinate roles — a *generator* whose output must survive deterministic symbolic extraction before any of it can become a fact, or a *renderer* forbidden by construction to add content. Neural components may suggest; they may never *be* the fact. Knowledge is therefore structured at ingestion time, so that answers are read from a typed store with source evidence attached, rather than sampled from next-token prediction at inference time. We report the system's current scale — approximately 372,764 structured knowledge records, 35 typed relation types (3 of them functional in the OWL sense), roughly 65,758 typed relation instances, and about 427,000 graph edges — and a from-scratch on-device model of approximately 18.7M parameters over a 5,919-word vocabulary. When generation is conditioned on Peel's symbolic structure, we measure a branch-diversity improvement from 0.589 to 0.966 on an internal metric. The system operates fully offline. We release a public dataset and a public on-device model. Peel is a PROTOTYPE and is not a clinical device; we state its limits plainly.

1 Introduction

1.1 The stakes

Augmentative and alternative communication (AAC) technology is used by people — many of them children — who cannot rely on natural speech. An AAC system does not merely *inform* its user;

it *speaks for* them. A statement the system surfaces can become the statement the user utters to a caregiver, a teacher, or a clinician.

This changes the cost of a hallucination. In a general-purpose chat assistant, a fabricated fact is an error the user can catch, discount, or verify. In AAC for a vulnerable user, a fabricated fact can be:

1. **Undetectable to the user**, who may be querying precisely because they do not already know the answer, and who may have limited capacity to challenge a confidently-stated falsehood.
2. **Laundered through the user's own voice**, so that a downstream listener attributes the fabrication to the person, not the tool.
3. **Consequential**, when the content concerns the user's own body, needs, medication, safety, or relationships.

We take as a design premise that in this domain a hallucinated "fact" is a safety failure, not a quality-of-service degradation. The appropriate engineering response is not to reduce the *probability* of fabrication but to remove the *structural pathway* by which fabrication can reach the user as fact.

1.2 Why statistical guardrails are insufficient

The dominant paradigm for reducing hallucination — better pretraining data, reinforcement learning from human feedback, retrieval augmentation, self-consistency, and post-hoc filtering — improves the *rate* at which a model produces false statements. Each of these techniques is valuable and none of them changes the underlying generative object: a probability distribution over next tokens. A system whose facts are the output of sampling has, by construction, a nonzero probability of emitting any syntactically valid falsehood. Lowering that probability is worthwhile; it is not the same guarantee as *a false statement cannot be authored as a fact.*

For a safety-critical, vulnerable-user domain we argue the field needs the second kind of guarantee — a **structural** property of the system's control flow — rather than a better point on the statistical curve.

1.3 Contributions

This paper makes the following contributions, each traceable to the Peel system as built:

1. **The inversion thesis (Section 3):** a labeled architectural principle in which a typed symbolic store is the sole author of facts and neural models are demoted to generator or renderer. We formalize the principle and its two consequences (generators must pass symbolic extraction; renderers may not add content).
2. **Structure-at-ingestion as an alternative to retrieval-at-inference (Section 4):** an account of *when* knowledge is structured — at entry, not at query time — and why this differs categorically from retrieval-augmented generation and from vector-store "memory."

3. **A property-level architecture for structural hallucination prevention (Section 5):** typed symbolic storage; provenance attached to every fact as an invariant; a deterministic contradiction check in which functional relations act as hard vetoes; bounded, transitive inference; and fully offline, on-device operation.

4. **Measured results at nontrivial scale (Section 6):** knowledge-base and model scale figures, a measured branch-diversity improvement under symbolic conditioning, offline operation, and two public artifacts (a dataset and an on-device model).

We deliberately describe the enforcement machinery at the level of *properties enforced*, not implementation recipes; the mechanisms are covered by a patent-sensitive internal document. This is the same discipline used when a system is specified by the invariants it guarantees rather than by the code that guarantees them.

2 Background and problem framing

2.1 AAC and the demand for correctness

AAC spans low-tech boards and high-tech speech-generating devices; a standard reference is Beukelman and Mirenda's text on the field [1]. What distinguishes the AAC setting for our purposes is the combination of (a) users who may not be positioned to verify what the system asserts and (b) output that is consumed as the user's own communication. These two facts together raise correctness from a usability concern to a safety concern.

2.2 Hallucination as a structural, not statistical, target

Hallucination in neural language models is a well-surveyed phenomenon [2]. The literature overwhelmingly frames both the problem and the solution statistically: measure the rate, reduce the rate. Our position is that for the AAC-for-vulnerable-users setting, the correct target is a control-flow invariant — *no output path exists by which an unextracted, unsupported statement becomes a stored or spoken fact* — and that such an invariant is achievable only if the neural component is prevented, by construction, from authoring facts at all.

3 The inversion thesis

We state the central principle as a named commitment.

> **The Inversion.** In a conventional neural assistant, a neural model *is* the source of the answer, and symbolic structures (if any) are auxiliary. Peel inverts this. A typed symbolic knowledge store is the **only author of facts.** Every neural model is demoted to exactly one of two subordinate roles:
>
> - a **generator**, whose output is untrusted proposal material and must survive deterministic symbolic extraction before *any part of it* can be admitted as a fact; or

- a **renderer**, which may reshape already-authored facts into fluent surface form but is **forbidden by construction from adding content**.

Neural components may *suggest*. They may never *be* the fact.

Two consequences follow directly and define the boundary that the rest of the architecture enforces:

- **C1 (generator subordination).** No neural output reaches the store as a fact except through a deterministic extraction step. What the extraction step cannot certify does not enter. The generator's fluency, plausibility, or confidence carries no authority.
- **C2 (renderer containment).** The component that produces the final surface text operates over facts the store has already authored. It is constrained so that content present in its output must trace to content present in its input. Fluency is permitted; invention is not.

The inversion is what converts hallucination from a statistical quantity to be minimized into a structural pathway that is closed. A generator can still *propose* a falsehood — that is inherent to any generator — but under C1 the proposal is inert until symbolic extraction certifies it, and under C2 nothing new can be smuggled in at the rendering stage. The fabrication has nowhere to become a fact.

4 Structure at ingestion, not retrieval at inference

The inversion implies a choice about *when* knowledge is organized.

Retrieval-at-inference — the pattern of retrieval-augmented generation (RAG) [3] and of vector-store "memory" — leaves knowledge as unstructured or semi-structured text and defers organization to query time. At a query, similar passages are retrieved by embedding proximity and handed to a generator, which composes an answer. The generator remains the author of the answer; retrieval only conditions it. Fabrication remains reachable, because the final synthesis step is still probabilistic and still free to interpolate beyond what was retrieved.

Structure-at-ingestion — Peel's choice — organizes knowledge *when it enters the system*. Incoming material is subjected to deterministic symbolic extraction; what survives becomes typed records in the store, each carrying its provenance. At query time the system reads from this typed store. The answer is *found*, not *composed*: it is a set of stored facts with source evidence attached, optionally passed to a renderer for fluency under C2.

The distinction is categorical, not a matter of degree. Under retrieval-at-inference, the authorship of the answer sits in a neural model at query time. Under structure-at-ingestion, authorship has already happened, deterministically, at entry — and the query path has no license to author anything new. This is why we regard Peel as offering a structural, rather than statistical, hallucination guarantee. A companion note, *Retrieval Is Not Memory* [4], develops the adjacent argument that retrieval is a proper subset of memory and that recall is not authorization to act; the present paper is the fact-authorship half of the same worldview.

5 Architecture as enforced properties

We describe Peel by the properties it enforces. Mechanism recipes are intentionally omitted (Section 10 notes the scrub).

5.1 Typed symbolic knowledge store

Facts live in a typed store rather than as free text or as opaque embeddings. Each record is described by a typed, multi-field schema spanning four families — core, semantic, metadata, and relational. Typing is load-bearing: it is what lets extraction certify a candidate as *a fact of a known kind*, what lets the contradiction check reason over relations, and what lets inference remain bounded. Relations between concepts are themselves typed; the current schema defines **35** relation types.

5.2 Provenance as an invariant

Every fact in the store is traceable to the source from which it was extracted. Provenance is not an optional annotation added after the fact; it is an admission condition. A candidate that cannot be tied to a source and its supporting evidence is not admitted. Consequently, any answer the system returns can be accompanied by *where it came from* — a property that matters especially when the consumer is a caregiver or clinician who may need to check the basis of what the user "said."

5.3 Deterministic contradiction checking with functional-relation vetoes

Before a candidate fact is admitted, it is checked for contradiction against what the store already holds. The check is deterministic. Among the 35 relation types, **3 are functional** in the OWL 2 `FunctionalProperty` sense [5] — meaning a subject may bear the relation to at most one object. Functional relations act as **hard vetoes**: a candidate that would give a subject a second, conflicting value for a functional relation is refused, not averaged, ranked, or probabilistically down-weighted. (The veto algorithm itself is part of the internal, patent-sensitive specification and is not disclosed here; what is public is the property: functional relations are enforced as at-most-one constraints that reject conflicting writes.)

Where the store lacks a relevant relation to reason over, the check **abstains** rather than guessing — returning an "unknown" verdict rather than a manufactured one. Abstention under ignorance is itself part of the honesty posture: the system declines to author a fact it cannot check.

5.4 Bounded, transitive inference

The system performs limited transitive inference over typed relations to answer questions not stored verbatim. Inference is bounded: derived conclusions are constrained in confidence and in scope, endpoints must be known concepts in the store, and derivation cannot manufacture concepts that were never ingested. The point is that inference *extends* the authored store along typed, checkable edges — it does not open a side channel for free-form generation.

5.5 Offline and on-device

Peel runs fully offline, on the device. There is no inference-time dependency on a remote model or a network service to author or check facts. For the target population this is a privacy property (a child's conversational material does not leave the device) and a reliability property (the system's correctness guarantees do not depend on connectivity). The accompanying small model is trained from scratch and sized to run on-device.

6 Results

The figures below are current measured characteristics of the system and its artifacts. They are results, not mechanisms.

6.1 Scale of the knowledge base and model

Table 1: Peel system scale (current).

Quantity	**Value**
Structured knowledge records	~372,764
Typed relation types defined	35
— of which functional (OWL `FunctionalProperty`)	3
Typed relation instances	~65,758
Graph edges	~427,000
On-device model parameters	~18.7M (18,667,008)
Model vocabulary (shipped)	5,919 words
Model training	from scratch, on-device-sized

6.2 Branch diversity under symbolic conditioning

We measure an internal branch-diversity metric under two generation regimes. When generation is conditioned on Peel's symbolic structure rather than on positional context alone, branch diversity rises from **0.589 to 0.966**. This is an internal metric, not a standard public benchmark, and we report it as such (Section 8). Its relevance is that symbolic conditioning materially broadens the generated distribution — evidence that the symbolic store contributes structure the neural component would not otherwise express.

6.3 Offline operation

The system authors, checks, and answers without network access. This is verified as an operating property of the on-device deployment.

6.4 Public artifacts

Two artifacts are public and downloadable:

- **Dataset:** `Hoodrobot/AAC_Children_Conversations` (Hugging Face).
- **On-device model:** `Hoodrobot/TinkyBrain-v3` (Hugging Face).

Table 2: Property-to-evidence map.

Enforced property	**Status**	**Evidence class**
Typed symbolic store is sole fact author	PROTOTYPE	System design; Table 1 scale
Provenance attached to every admitted fact	PROTOTYPE	Admission-condition property
Contradiction check deterministic; functional relations veto	PROTOTYPE	3 functional relations of 35; abstain-on-unknown
Bounded transitive inference	PROTOTYPE	Confidence/scope bounds; known-concept endpoints
Offline / on-device	PROTOTYPE	Operating property; on-device model
Symbolic conditioning improves branch diversity	Measured (internal metric)	$0.589 \rightarrow 0.966$
Public dataset + model released	Released	HF artifacts above

7 Why this matters for safety-critical AI and vulnerable users

The AAC setting is an unusually clean instance of a general problem: when an AI system's output is consumed as ground truth by a party who cannot independently verify it, statistical hallucination-reduction is the wrong guarantee. The right guarantee is structural — a property of the system that holds regardless of what the generator happens to sample.

Peel's inversion is one way to obtain that property. By making a typed symbolic store the only author of facts, attaching provenance as an admission condition, and enforcing functional-relation contradictions as hard vetoes, the system converts "we hope the model rarely lies" into "the model cannot author a fact the store did not certify." We do not claim this is the only route to structural hallucination prevention, nor that Peel's guarantees are complete (Section 8). We claim that the *shape* of the guarantee — structural, deterministic, provenance-bearing — is the shape the vulnerable-user domain requires, and that Peel demonstrates it is buildable and runs offline at nontrivial scale.

8 Limitations

We label Peel honestly. It is a **PROTOTYPE**. It is **not a clinical device**, has not undergone clinical validation, and must not be represented as a medical or diagnostic instrument. The following limitations bound every claim above.

1. **Prototype maturity.** Peel is a research prototype, not a production or enterprise system. It lacks the full reliability, security, and operational envelope required to ship to vulnerable users unsupervised.

2. **The guarantee is about fact authorship, not truth.** The inversion prevents *unextracted, unsupported* statements from becoming stored facts. It does not guarantee that ingested source material is itself correct: a false fact from a trusted-but-wrong source, correctly extracted and provenanced, can still enter the store. Peel closes the *fabrication* pathway; it does not adjudicate the *veracity* of well-formed sources. Source curation is a separate, unsolved responsibility.

3. **Extraction coverage is not completeness.** Deterministic symbolic extraction admits only what it can certify. This is intentional (it is what makes the guarantee hold), but it means valid content that the extractor cannot yet handle is dropped rather than admitted. Coverage of the extraction step is a bounded, ongoing quantity, not a solved one.

4. **The diversity result uses an internal metric.** The $0.589 \rightarrow 0.966$ branch-diversity improvement is measured on an internal metric, not a standard public benchmark, and has not been externally replicated. It should be read as directional evidence of symbolic conditioning's effect, not as a benchmarked accuracy claim.

5. **No end-to-end clinical or user-study evaluation is reported here.** We report system scale, an internal generation metric, offline operation, and public artifacts. We do not report a controlled study with AAC users measuring communication outcomes or real-world hallucination incidence. That study does not yet exist and is required before any clinical claim.

6. **Contradiction checking is only as strong as the relation schema.** Only relations that are typed — and, for hard vetoes, only those that are functional — are enforced. Contradictions that are not expressible in the current 35-relation schema are not caught by the check; they may pass as "unknown." Expanding the schema expands coverage but the schema is finite and incomplete.

7. **Abstention is a feature, not a limitation, but it has a cost.** The system's willingness to answer "unknown" rather than guess reduces fabrication but also reduces coverage; some answerable-in-principle questions will be declined.

8. **Mechanism scrub.** This public document withholds the enforcement mechanisms (extraction, veto algorithm, provenance schema internals, and diversity-conditioning method). Readers cannot independently reimplement the guarantees from this document alone. That is a deliberate constraint of a moat-scrubbed release, and it limits external reproducibility here.

9 Conclusion

Peel is built on a single refusal: the refusal to let a probability distribution author a fact for a child who cannot check it. By inverting the usual arrangement — a typed symbolic store as the only author of facts, neural models demoted to generator or renderer, knowledge structured at ingestion and provenanced on admission, contradictions vetoed deterministically by functional relations, and the whole system run offline — we obtain a structural rather than statistical stance toward hallucination. The prototype demonstrates that this stance is buildable at nontrivial scale and runs on-device. Much remains unproven, and we have said so. But the design commitment is the point, and it reduces to one line: **facts should be authored by structure, not sampled by probability.**

10 Note on scope and scrub

This document is the public, moat-scrubbed counterpart to a patent-sensitive internal whitepaper. It conveys the thesis, the problem, the property-level architecture, and the measured results. It deliberately omits the enforcement mechanism recipes — the extraction pipeline, the contradiction-veto algorithm, the diversity-conditioning method, and the provenance schema internals — which are the subject of the internal document. The convention throughout is to specify Peel by the properties it enforces rather than the implementation that enforces them.

References

[1] D. R. Beukelman and P. Mirenda. *Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs.* 4th ed., Brookes Publishing (ISBN 978-1-59857-196-7). Standard AAC reference.

[2] Z. Ji, N. Lee, R. Frieske, et al. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55(12), Article 248, 2023. doi:10.1145/3571730 (arXiv:2202.03629).

[3] P. Lewis, E. Perez, A. Piktus, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *NeurIPS*, 2020. arXiv:2005.11401.

[4] Perslis Research. "Retrieval Is Not Memory: Memory as a Governance Function over Experience." 2026. `https://research.perslis.com/memory.html`

[5] Perslis Research. "The Orchestration Gap." 2026. `https://research.perslis.com/orchestration-gap.html`

[6] W3C. *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax — `FunctionalProperty`*. W3C Recommendation. `https://www.w3.org/TR/owl2-syntax/`

Public artifacts.
Dataset: `Hoodrobot/AAC_Children_Conversations` (Hugging Face).
Model: `Hoodrobot/TinkyBrain-v3` (Hugging Face).
