Gentags: Discrete Semantic State for Constraint-Sensitive Decision Pipelines

Anonymous ACL Submission

Abstract

LLM pipelines that act on textual evidence face two practical difficulties: (1) source text often exceeds the context window, requiring compression, and (2) even when meaning is present in context, it remains embedded in free-form prose, making it difficult to diagnose, reference, or attribute to specific decisions. We introduce **Gentags**, a representation that compresses source text into short, evidence-grounded semantic units, each isolating one identifiable semantic condition. To evaluate whether this representation preserves decision-relevant meaning, we compare Gentags against lexical baselines (RAKE, YAKE, TF-IDF) in a constraint-sensitive decision setting that isolates representation structure while holding evidence, task, and evaluation fixed. Gentags improve agreement with full-evidence decisions to **79.5%** (vs. **52.3–61.6%**) and raise hard-constraint satisfaction to **97.3%** (vs. **84.7–89.3%**), while showing greater stability across runs, prompts, and extractor models. These results suggest that discrete semantic representations can effectively preserve decision-relevant meaning, and motivate further exploration of discrete semantic compression as intermediate state in LLM pipelines.

1 Introduction

LLM pipelines that process textual evidence routinely compress source text before downstream processing, through chunking, retrieval, summarization, or keyword extraction, because the full evidence often exceeds what fits in a single context window (Lewis et al., 2020). Recent work has shown that retrieval and chunking strategies remain critical determinants of answer quality even as context windows grow (Li et al., 2024), and improvements continue to focus on chunk boundaries, contextual enrichment, and retrieval granularity, with finer-grained units such as propositions outperforming passage-level chunks (Chen et al., 2024). Across all of these approaches, however, the intermediate representation passed to downstream models remains free-form text—passages, summaries, or extracted fragments organized by document structure, token limits, or retrieval scores, not by semantic content.

This means that decision-relevant meaning within these representations remains difficult to decompose. A passage may express multiple semantic conditions implicitly, redundantly, or across sentence boundaries. Even when the right text is retrieved, the specific conditions that matter for a given task are not separated from the surrounding narrative—making it difficult to diagnose which semantic factors are present, reference them individually, or attribute downstream outcomes to specific pieces of meaning. This is a representational problem, not only a retrieval or scaling problem.

Prior work supports the premise that representation structure matters. Intermediate representations have been shown to materially affect downstream correctness through reasoning traces, problem decomposition, multi-path aggregation, and executable intermediates (Wei et al., 2022; Zhou et al., 2022; Wang et al., 2023; Gao et al., 2023; Yao et al., 2023). Work on semantic parsing and open information extraction has demonstrated that structured meaning can be derived from text (Zettlemoyer and Collins, 2005; Banko et al., 2007), and systems such as MemGPT have shown that externalizing information into persistent structured memory can change what a model is able to do across interactions (Packer et al., 2023). Yet these lines of work either treat extracted representations as evaluation endpoints, decompose model outputs for verification, or operate over free-form text rather

than explicit semantic units. What remains open is whether compressing source evidence into discrete semantic units can function as an effective intermediate representation for downstream procedures.

We explore that question by introducing **Gentags**: short, evidence-grounded semantic units that compress source text into discrete, individually addressable units that fit within the context window of a downstream model. Rather than improving how text is chunked or retrieved, Gentags represent a different axis: the source text is decomposed by semantic content rather than by document structure or token boundaries. Each unit isolates one identifiable semantic condition supported by the evidence.

The central comparison is between Gentags and lexical baselines—RAKE, YAKE, and TF-IDF—that also produce discrete units from the same source text. These baselines are a natural reference point: they are widely used for text compression in information retrieval practice, they are inexpensive to obtain, and they produce units that can be passed directly into a context window. This design isolates representation structure: the source evidence, downstream task, judge model, and evaluation procedure are held fixed, while only the intermediate representation changes.

To provide a rigorous evaluation setting, we use constraint-sensitive decision-making as a testbed—a setting where explicit requirements must be checked against evidence, producing binary, auditable outcomes. We evaluate across three dimensions: stability across runs, prompts, and extractor models; structural properties of the resulting representation; and downstream decision fidelity in a setting involving 50 venues, four personas with explicit hard requirements, and majority-vote aggregation over five repeated judge calls with two independent judge models.

This work makes three contributions:

1. We frame the structure of intermediate semantic representations as a design variable in LLM-based evidence pipelines.

2. We introduce Gentags as a discrete, evidence-grounded representation that compresses source text into individually addressable semantic units.

3. We provide controlled empirical evidence that propositional representations improve decision reliability relative to fragment-level lexical representations in this setting, including under token-matched conditions that control for information volume.

2 Related Work

2.1 Structured semantic representations from language

A long line of work has studied how to derive structured representations of meaning from natural language. Semantic parsing maps text into formal representations such as logical forms or executable programs (Zettlemoyer and Collins, 2005; Dong and Lapata, 2016). Open information extraction systems extract relational units from text without requiring a fixed schema (Banko et al., 2007). These approaches demonstrate that structured meaning can be derived from text, but they typically rely on predefined output schemas or relational ontologies and evaluate extracted representations as endpoints using precision, recall, and F1 against gold-standard annotations.

2.2 Aspect-based sentiment analysis

Aspect-based sentiment analysis (ABSA) extracts discrete aspect-level units from text and associates them with sentiment polarities (He et al., 2019; Wang et al., 2016). These units are short, discrete, and derived from evidence text. However, ABSA assumes predefined aspect and sentiment dimensions and treats extraction as the endpoint, evaluated by F1 against annotated gold standards rather than by downstream task performance.

2.3 Claim decomposition and semantic unit extraction

Recent work on hallucination detection and factuality verification decomposes generated text into discrete semantic units for downstream evaluation. RefChecker extracts claim-triplets from LLM outputs and evaluates each against a reference to detect fine-grained hallucinations (Hu et al., 2024). FacTool decomposes generated content into verifiable units across domains to identify factual errors (Chern et al.,

2023). ERASER provides rationale extraction benchmarks that evaluate which parts of input text support downstream decisions (DeYoung et al., 2020). These systems treat semantic units as operational interfaces for downstream judgments, but they decompose model outputs or generated claims for verification against ground truth rather than compressing source evidence into an intermediate decision state.

2.4 Intermediate representations in LLM pipelines

Intermediate representation structure can materially affect downstream correctness and task success—through reasoning traces, problem decomposition, multi-path aggregation, and executable intermediates (Wei et al., 2022; Zhou et al., 2022; Wang et al., 2023; Gao et al., 2023; Yao et al., 2023). These approaches introduce structure into reasoning trajectories, tool interactions, or retrieval contexts, but the semantic content itself typically remains embedded in free-form text spans rather than explicitly externalized as discrete units.

Separately, retrieval-augmented generation provides supporting evidence during generation by retrieving relevant passages from external sources (Lewis et al., 2020). Recent work has shown that retrieval and chunking strategies remain critical determinants of answer quality even as context windows grow (Li et al., 2024), and that finer-grained retrieval units such as propositions outperform passage-level chunks (Chen et al., 2024). These advances improve *which* text reaches the model and at *what granularity*, but the intermediate representation remains free-form text—the decomposition serves retrieval rather than producing an explicit semantic representation for downstream procedures. Systems like MemGPT externalize information into persistent structured memory to support reuse across interactions (Packer et al., 2023), but the resulting memory streams are not discrete, evidence-conditioned semantic units designed to be consumed as an intermediate representation.

2.5 Positioning

Across these lines of work, structured meaning extraction evaluates representations as endpoints against human-labeled annotations; claim decomposition targets model outputs for verification; retrieval improvements refine which text reaches the model and at what granularity; and intermediate reasoning structures operate over free-form text rather than explicit semantic units. None directly study whether compressing source evidence into discrete semantic units can function as an effective intermediate representation for downstream procedures. We investigate that question, using constraint-sensitive decision-making as a controlled evaluation setting.

3 Gentags

A gentag is a short, schema-free semantic unit grounded in source text. For a venue with evidence E, extraction produces a set $S = f_\theta(E) = \{g_1, \ldots, g_n\}$ whose members express properties, attributes, or entities supported by the reviews rather than verified world facts. Because Gentags are discrete natural-language units, they can be inspected, compared across runs, cached, and evaluated directly against decision constraints.

3.1 Extraction

Gentags are produced by prompting a large language model with textual evidence and requesting short gentags grounded in the input text. No predefined label set or ontology is supplied; the model generates gentags freely based on the evidence.

Given textual evidence E, extraction produces a gentag state

$$S = f_\theta(E)$$

where f_θ denotes the extraction model parameterized by θ. The resulting state

$$S = \{g_1, g_2, \ldots, g_n\}$$

is an unordered set of gentags representing properties, attributes, or entities supported by the evidence. The number of gentags n varies depending on the information contained in the source text.

Extraction is performed in a zero-shot setting, meaning that the model receives only the evidence and instructions to produce short gentags grounded in the text. Implementation details, including prompts and model configurations, are provided in the appendix.

3.2 Gentag State Representation

The gentag state of an entity is represented as a finite set of gentags:

$$S = \{g_1, g_2, \ldots, g_n\}.$$

Each gentag g_i is a short natural-language string representing a property, attribute, or entity supported by the evidence. The set S therefore forms an explicit semantic state abstraction derived from the source text.

Unlike dense embeddings, which encode meaning in continuous vector spaces, and unlike raw text, which embeds meaning in unstructured narrative, the Gentags representation exposes properties, attributes, or entities as individually addressable gentags.

3.3 Representation Properties

The gentag representation has several properties that distinguish it from alternative text-derived representations.

Discrete. Each gentag is a separable unit, and the state is a finite set of bounded strings rather than a continuous vector or paragraph of text.

Schema-free. Gentags are not drawn from a predefined taxonomy or label set. The vocabulary emerges from the interaction between the language model and the evidence.

Externalized. The semantic state is stored outside the language model as persistent data, allowing it to be cached, compared, and reused independently of the extraction model.

Evidence-conditioned. Each gentag is grounded in the provided source text. The representation therefore reflects properties, attributes, or entities supported by the evidence rather than arbitrary model-generated descriptions.

Inspectable. Because gentags are natural-language strings, the resulting state can be directly read and interpreted by humans without additional decoding or projection.

4 Experimental Setup

We evaluate Gentags through three complementary studies targeting different properties of the representation: **stability**, **structural organization**, and **downstream decision utility**. Across all studies, comparisons are performed at the representation level, and the underlying textual evidence and evaluation protocols are held fixed wherever possible.

4.1 Data

All studies draw from a shared corpus of 553 venues, each associated with 1–20 user reviews collected via the Google Maps API. Gentags and all baseline representations are derived from this same underlying evidence.

Different studies operate on fixed subsets of this corpus depending on their requirements. Stability and structural analyses use a subset of venues with successful extractions across all models to enable controlled comparison. The decision evaluation uses a stratified subset of 50 venues selected to ensure balanced presence and absence of sports-viewing and fast-service indicators to avoid floor/ceiling effects on compliance.

4.2 Representations and Baselines

The primary object of study is the **gentag state**, defined as a set of short, evidence-conditioned semantic units extracted from text.

We compare Gentags to several text-derived baseline representations constructed from the same review evidence:

- **RAKE**: keyword phrases extracted using Rapid Automatic Keyword Extraction (Rose et al., 2010).
- **YAKE**: unsupervised keyphrase extraction based on local features (Campos et al., 2020).
- **TF-IDF**: top-weighted n-grams or phrases (Salton and Buckley, 1988).
- **gentag_truncated**: Gentags truncated to match the number of RAKE phrases per instance.
- **FER (Full-Evidence Reference)**: raw review text provided directly to the decision model, serving as a practical upper bound—the best available reference given the absence of human ground-truth decisions.

All comparisons are **representation-to-representation**: the underlying evidence, prompts, and evaluation procedures are held constant, and only the representation of semantic content varies.

The `gentag_truncated` condition controls for representation size. If truncated Gentags outperform lexical baselines under equal-length conditions, performance differences can be attributed to semantic structure rather than information volume.

One design asymmetry should be noted: Gentags are produced by LLM-based extraction, while the lexical baselines (RAKE, YAKE, TF-IDF) are statistical methods. The truncated condition controls for information volume but not for generation method. It is therefore possible that part of the observed advantage reflects the use of an LLM extractor rather than the specific representational structure of Gentags. Evaluating Gentags against stronger semantic baselines—such as LLM-generated summaries or aspect-level extractions—is left to future work.

4.3 Study Design

Stability Analysis. The stability study evaluates whether Gentags behave as a recoverable semantic representation under variation in extraction conditions. Extractions are performed across four language models — OpenAI (gpt-5-nano), Google (gemini-2.5-flash), Anthropic (claude-sonnet-4-5), and xAI (grok-4) — across three prompt variants and repeated runs per model-prompt combination. Stability is assessed by comparing outputs for the same input under these variations.

Structural Analysis. The structural analysis evaluates how semantic content is organized within each representation. Tags and baseline phrases are mapped to a fixed set of diagnostic semantic facets using embedding-based similarity. This enables measurement of how concentrated or dispersed semantic information is across interpretable dimensions.

All representations are evaluated under the same facet definitions, embedding model, and assignment procedure. Coverage and concentration results are relative to this specific diagnostic probe space, not absolute measures of semantic breadth.

Decision Evaluation. The decision study evaluates whether representations preserve decision-relevant information in a constraint-sensitive setting. Each condition consists of a venue, a persona, and a representation. A model receives only the representation and produces a decision.

We evaluate 50 venues, 4 personas, and 6 representation conditions, with repeated model evaluations per condition and majority-vote aggregation.

4.4 Personas and Constraints

The decision evaluation uses four personas. Three impose explicit hard constraints, and one serves as a soft-preference control:

- **Food Critic**: must reject venues with negative food-quality indicators.
- **Sports Fan**: must reject venues lacking sports-viewing indicators.
- **Quick Lunch Worker**: must reject venues lacking fast-service indicators.
- **Balanced Diner**: no hard constraints.

Constraint evaluation is based on fixed indicator lexicons. These lexicons are held constant across all representations and are not tuned post hoc.

4.5 Evaluation Protocol

Decision evaluation is performed using language models acting as judges. The primary judge is GPT-4o (`gpt-4o-2024-08-06`), and results are validated using Claude Sonnet 4 (`claude-sonnet-4-20250514`) as a second independent judge.

Each condition is evaluated with $N = 5$ repeated judge calls, and final decisions are obtained via majority-vote aggregation. Invalid outputs are discarded, and conditions with insufficient valid responses are excluded from analysis.

For structured representations, the judge is required to return decisions along with explicit supporting and blocking evidence drawn only from the provided representation. Validation enforces that all cited evidence must be a subset of the input representation.

4.6 Controlled Factors

The experimental design isolates the effect of representation structure by controlling other variables:

- **Fixed evidence**: all representations for a given instance are derived from the same underlying text.
- **Fixed evaluation protocol**: prompts, decision criteria, and aggregation procedures are identical

Table 1: Run-to-run stability summary across repeated gentag extractions.

Metric	Median	Q1	Q3
Semantic cosine	**0.977**	0.968	0.986
Surface Jaccard	**0.471**	0.333	0.625
Mean Max Cosine	**0.887**	0.839	0.927
Semantic–surface gap	**0.504**	—	—

across representations.

- **Frozen constraints**: persona definitions and indicator lexicons are fixed in advance.
- **Controlled representation size**: the truncated Gentag condition matches baseline length.
- **Cross-model validation**: decision evaluation is repeated with an independent judge model.

Under this design, differences in downstream performance are attributable to variation in how representations preserve and expose decision-relevant semantic content, subject to the generation-method confound noted in §4.2.

5 Analysis / Results

We report results across three layers: stability (§5.1), structural organization (§5.2), and downstream decision utility (§5.3).

5.1 Stability Analysis

The stability results ask whether Gentags behave like a recoverable semantic state rather than a brittle surface artifact.

5.1.1 Run-to-run Stability

Across repeated extractions under the same model and prompt, Gentags show very high semantic consistency despite substantial lexical variation.

Semantic cosine is extremely high, while surface overlap is much lower. The gap shows that variation lies in paraphrase and lexical choice rather than semantic drift (Figure 1). This pattern holds across all four extractor models (Table 2).

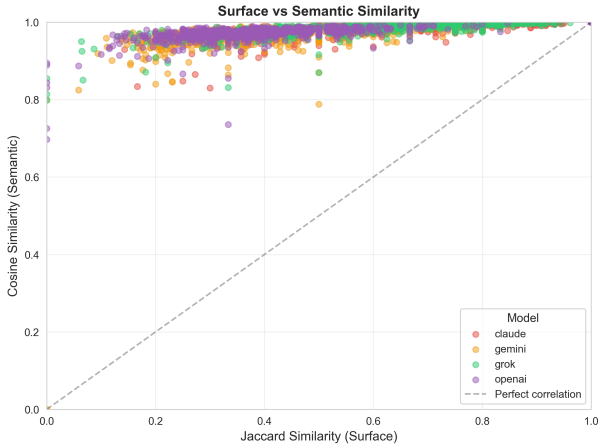

Figure 1: Surface vs. semantic decoupling. Points cluster in the upper-left region, indicating high semantic cosine despite lower surface Jaccard.

Table 2: Run-to-run stability metrics by extractor model.

Model	Cosine	Jaccard	Mean Max Cosine
Claude	0.982	0.574	0.913
Gemini	0.971	0.404	0.869
Grok	0.975	0.722	0.876
OpenAI	0.975	0.387	0.861

5.1.2 Prompt Sensitivity

Prompt variation affects phrasing and granularity but not core semantic content. All cross-prompt cosine values remain above 0.95 (Table 3).

Table 3: Cross-prompt stability across gentag extraction prompt variants.

Prompt Pair	Mean Cosine	Mean Jaccard
`anti_hallucination` ↔ `minimal`	0.966	0.321
`anti_hallucination` ↔ `short_phrase`	0.962	0.282
`minimal` ↔ `short_phrase`	0.966	0.352

5.1.3 Cross-model Agreement

Semantic agreement remains high across extractor models. All pairs exceed 0.94 cosine similarity (Table 4), indicating recovery of a shared semantic state.

Variability also decreases with evidence availability (Pearson $r = -0.230$, $p < 0.001$), suggesting that dispersion reflects identifiability rather than arbitrary noise (Appendix C.1.1).

Table 4: Cross-model stability across extractor pairs.

Model Pair	Mean Cosine	Mean Jaccard
Claude ↔ Gemini	0.951	0.253
Claude ↔ Grok	0.953	0.267
Claude ↔ OpenAI	0.951	0.236
Gemini ↔ Grok	0.969	0.323
Gemini ↔ OpenAI	0.958	0.248
Grok ↔ OpenAI	0.969	0.315

5.2 Structural Analysis

This analysis examines how Gentags organize semantic content relative to lexical baselines.

5.2.1 Facet Coverage and State-Gini

Gentags place more semantic mass into the diagnostic facet space and distribute it more evenly across facets.

Table 5: Facet coverage and assignment rates by representation.

Method	Mean tags	Assigned mean	Other mean	Other rate
Gentags	21.9	12.3	9.6	∼43%
RAKE	19.5	6.1	13.3	∼68%
TF-IDF	19.8	6.6	13.2	∼67%
YAKE	19.8	6.5	13.3	∼67%

Gentags show lower `other_rate` and lower State-Gini, whereas baselines show higher `other_rate` and higher Gini. The correct interpretation is joint: Gentags recover more decision-relevant units *and* distribute them more broadly across facets.

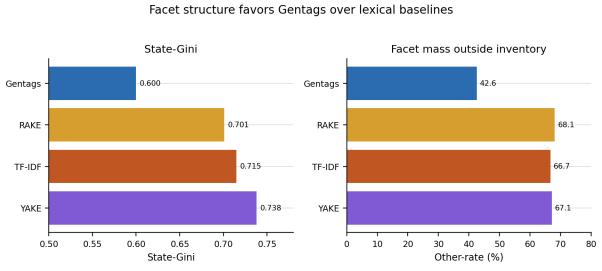

Figure 2: Joint view of State-Gini and facet coverage, showing broader and more balanced gentag structure.

5.2.2 Threshold Sensitivity

To test robustness, we reran facet assignment at $\tau \in \{0.30, 0.35, 0.40\}$. At every threshold, Gentags retain more mass within the facet inventory than lexical baselines, and the gap widens as τ increases (Table 6). A separate bleed check confirms this pattern is not driven by unusually noisy facet assignments (Appendix C.2.1).

Table 6: Threshold sensitivity of State-Gini and other-rate across methods.

τ	Method	State-Gini	Other rate (%)
0.30	Gentags	0.575	26.6
0.30	RAKE	0.647	50.9
0.30	TF-IDF	0.689	50.5
0.30	YAKE	0.710	50.2
0.35	Gentags	0.600	42.6
0.35	RAKE	0.701	68.1
0.35	TF-IDF	0.715	66.7
0.35	YAKE	0.738	67.1
0.40	Gentags	0.630	55.1
0.40	RAKE	0.733	78.0
0.40	TF-IDF	0.774	80.2
0.40	YAKE	0.770	79.6

5.3 Decision Evaluation

The decision evaluation uses 50 venues, 4 personas, and 6 systems, yielding 1,200 conditions per judge. Each condition is evaluated with majority vote over $N = 5$ repeated judge calls.

5.3.1 FER Agreement

The primary fidelity metric is agreement with Full-Evidence Reference (FER) decisions—the same judge, rubric, and aggregation procedure applied to raw reviews rather than compressed representations.

Table 7: FER agreement by representation.

System	Matches	Total	Agreement	Kappa
Gentag	159	200	79.5%	0.667
RAKE	122	198	61.6%	0.388
YAKE	117	200	58.5%	0.351
TF-IDF	104	199	52.3%	0.258
Gentag truncated	149	199	74.9%	0.596

These are large effects: Gentags exceed RAKE by nearly 18 points, YAKE by 21 points, and TF-IDF by over 27 points. Kappa values show the same pattern, with Gentags in the substantial-agreement range and lexical baselines in fair-agreement ranges.

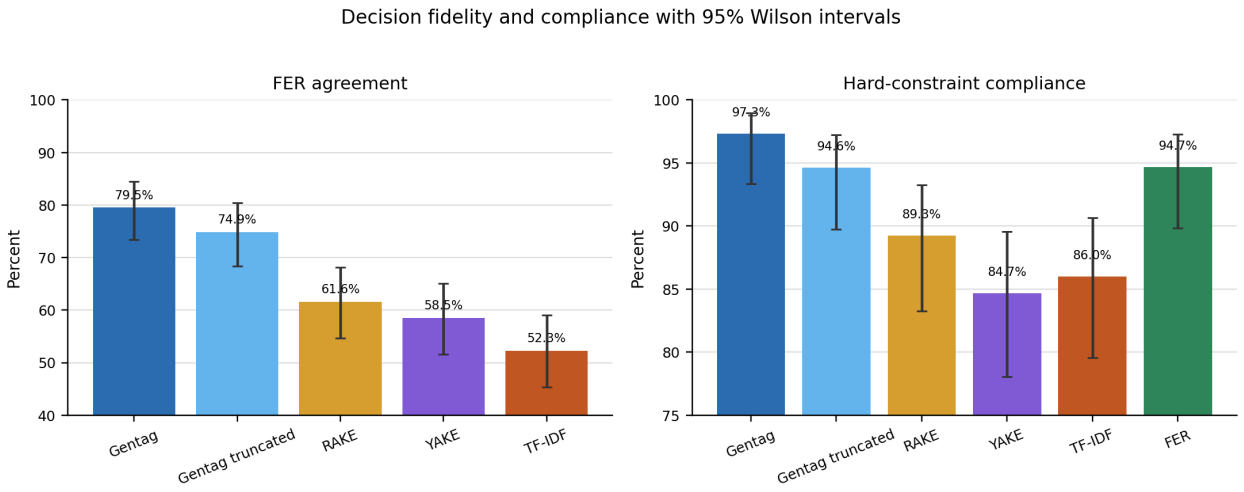

Figure 3: FER agreement comparison across representations, highlighting gentag gains over lexical baselines.

Table 8: Pairwise significance tests for FER agreement against gentags.

Comparison	p-value
gentag vs RAKE	0.0001
gentag vs YAKE	0.000008
gentag vs TF-IDF	< 0.0001

The strongest separation appears for the Sports Fan and especially the Quick Lunch Worker personas, where the Food Critic functions as a control with no negative food indicators in this sample.

5.3.2 Constraint Compliance

Hard-constraint compliance directly measures whether a representation enables correct decisions under explicit persona requirements. The same per-persona pattern holds (Table 10).

Table 9: Hard-constraint compliance by system.

System	Correct	Total	Compliance
Gentag	146	150	97.3%
FER	142	150	94.7%
Gentag truncated	141	149	94.6%
RAKE	133	149	89.3%
TF-IDF	129	150	86.0%
YAKE	127	150	84.7%

The Quick Lunch Worker is the clearest case: gentag phrases such as `"fast service"` and `"quick counter service"` communicate the speed constraint directly, whereas lexical fragments such as `"relative quick time"` are less decision-legible. The representation can fail to carry the exact constraint signal needed by the judge.

Table 10: Hard-constraint compliance by persona for the indicator-sensitive personas.

Persona	Gentag	RAKE	YAKE	TF-IDF
Sports Fan	96.0%	91.8%	88.0%	88.0%
Quick Lunch Worker	96.0%	76.0%	66.0%	70.0%

5.3.3 Token-budget Ablation

One possible objection is that Gentags perform better simply because they carry more information. The token-budget ablation addresses this by truncating Gentags to match the RAKE tag count per venue.

Table 11: Token-budget ablation under matched tag counts.

Metric	Gentag truncated	RAKE	YAKE	TF-IDF
FER agreement	74.9%	61.6%	58.5%	52.3%
Combined compliance	94.6%	89.3%	84.7%	86.0%

Matching information budget does not erase the gentag advantage. The ablation therefore isolates a representational effect: Gentags work better because of what they encode, not only how much they encode.

5.3.4 Cross-judge Agreement

To test robustness to evaluator choice, the entire decision study was rerun with a second judge model

(Table 12). An overall $\kappa = 0.712$ indicates substantial agreement, and the main decision pattern is not specific to a single judge model. Decision entropy analysis further confirms that lexical baselines systematically shift probability mass toward `BORDERLINE`, producing less decisive downstream behavior (Appendix C.3.1).

Table 12: Cross-judge agreement by system.

System	Matches	Total	Agreement	Kappa
FER	167	199	83.9%	0.731
Gentag	147	176	83.5%	0.746
Gentag truncated	144	173	83.2%	0.744
RAKE	160	191	83.8%	0.744
TF-IDF	135	178	75.8%	0.643
YAKE	137	177	77.4%	0.660
Overall	890	1094	81.3%	0.712

5.4 Overall Interpretation

Across all three layers, the same pattern emerges. Gentags are semantically stable across reruns, prompts, and models. They produce broader and more balanced facet coverage than lexical baselines, and this structural advantage transfers downstream: higher agreement with full-evidence decisions, more reliable constraint satisfaction, and more decisive output distributions—including under token-matched conditions.

These results suggest that discrete semantic representations can effectively preserve decision-relevant meaning, and motivate further exploration of discrete semantic compression as intermediate state in LLM pipelines.

6 Discussion

The main takeaway of this paper is not only that Gentags outperform lexical baselines in this setting. It is that the structure of the intermediate representation changes what kind of semantic information a pipeline can carry forward. Once meaning is externalized as discrete, addressable units that fit within a context window, compression does not only remove information—it creates an explicit representation that can be inspected, compared across runs, and consumed more legibly by downstream procedures.

One observation worth highlighting is that Gentag constraint compliance (97.3%) slightly exceeds the full-evidence reference condition (94.7%). Because FER is a practical upper bound rather than ground truth, this gap should not be read as evidence that Gentags are more correct than full evidence. The more likely explanation is signal-to-noise: full review text includes hedging, contradictory cues, and irrelevant material that can push the judge away from the constraint-relevant signal, while Gentags filter that content into discrete units that are more directly legible under an explicit constraint-checking rubric. This is an observation about the interaction between representation structure and judge behavior in this setting, subject to the generation-method confound noted in §4.2, not a general claim about the superiority of compressed representations over full evidence.

6.1 What this opens up

Once semantic units are explicit, a natural next step is conflict-aware source aggregation. If different sources support competing propositions, future systems could aggregate evidence at the semantic level rather than leaving disagreement buried in raw text. Gentags already make those conflicts visible; richer future variants could make them computable. A related direction is evidence-sensitive updating: if Gentags carried support counts, provenance, or confidence-like signals, the state could be revised as new evidence arrives rather than simply accumulating more tags—moving from a one-shot extraction layer toward a reusable semantic substrate.

More broadly, explicit semantic representations open the door to reasoning over addressable units rather than repeatedly re-reading free-form text. That could support rule-based systems, LLM judges, hybrid pipelines, and agentic systems that maintain reusable semantic memory across steps. The present paper does not demonstrate those broader systems, but it shows why explicit semantic representations make them more tractable.

7 Limitations

Scope. The evaluation is confined to a single restaurant-review decision domain. This domain permits controlled hard constraints, repeated extraction, and manual inspection of disagreements, but it is narrow. The results show that Gentags improve decision

fidelity in this setting, not that discrete semantic representations are universally preferable across tasks or domains.

Generator-class confound. Gentags are produced by a language model, whereas the lexical baselines (RAKE, YAKE, TF-IDF) are statistical algorithms with no semantic interpretation. The observed advantage therefore cannot be attributed solely to discrete semantic organization independent of LLM-based generation. Evaluating against stronger semantic baselines—including alternative LLM-generated discrete representations with different structural properties—is an important direction for future work.

Representation gaps. Current Gentags record what semantic conditions are present, but not how strongly each is supported, how uncertain it is, or how contradictory evidence should be interpreted. The dominant failure mode in the audit—mixed-evidence drift into `BORDERLINE` (33/41 mismatches)—suggests that Gentags often preserve relevant propositions but lack the evidential metadata needed to resolve ambiguity. Canonicalization is also incomplete: semantically equivalent tags may still appear as separate units rather than being merged, which will matter more for systems that aggregate across time or repeated extractions.

Evaluation artifacts. The structural analysis is probe-relative: facet coverage and State-Gini are measured against a specific 10-facet diagnostic space and should be interpreted accordingly. Part of the remaining decision error is introduced by evaluation protocol artifacts rather than representational failure—four of eight full reversals arise from frozen exact-match indicator lexicons where the Gentag state contains semantically relevant signals that are not recognized. Extraction reliability also remains model-dependent, with schema-adherence failures reducing effective sample sizes for some extractor models. Finally, the downstream evaluation relies on LLM judges rather than human adjudication. Human evaluation of this scale—1,200 conditions per judge with majority-vote aggregation—would be prohibitively costly (Min et al., 2023), and prior work has shown that automated evaluation can closely approximate human judgments in similar settings. Matched protocols, cross-judge replication, and the qualitative audit reduce this concern but do not eliminate it.

8 Conclusion

We introduced Gentags, a representation that compresses source text into discrete, evidence-grounded units, and evaluated whether this representation preserves decision-relevant meaning in a controlled constraint-sensitive setting. The results show that Gentags are recoverable across models and prompts, provide broader coverage of decision-relevant facets than lexical baselines, and improve both agreement with full-evidence decisions and hard-constraint compliance—including under token-matched conditions. These results suggest that discrete compressed representations can effectively preserve decision-relevant meaning, and motivate further exploration of this approach as intermediate state in LLM pipelines. Code and data will be made available upon publication.

9 Ethics Statement

The review corpus is derived from Google Maps platform data accessed via the Google Maps API, used in accordance with Google's Terms of Service. Raw data is not released with this submission in compliance with those terms. Derived experimental artifacts—extracted representations, decision outputs, and analysis tables—do not contain personally identifiable information. Venue names appear in qualitative examples for interpretability purposes only.

Extraction and evaluation rely on commercial LLM APIs. These models may reflect biases in how they interpret evidence, apply constraints, or resolve ambiguity. A representation that appears interpretable is not automatically fair or unbiased, and model-specific behavior can influence which semantic content survives compression.

Externalized semantic state can improve auditability relative to opaque free-text pipelines, but can also create a misleading sense of transparency. Gentags are evidence-conditioned summaries derived from text, not verified facts about described entities.

This work studies representations in a controlled

domain with synthetic personas and should not be taken as evidence that this pipeline is appropriate for consequential decisions in domains such as hiring, credit, healthcare, or legal adjudication. Any such extension would require domain-specific validation, human oversight, and fairness evaluation well beyond what is provided here.

10 Reproducibility and Materials

The paper is backed by versioned scripts, frozen analysis subsets, run-level outputs, and paper-facing documentation. We release extraction prompts, evaluation prompts, frozen persona definitions, indicator lexicons, analysis code, and the full set of extracted Gentags via the accompanying repository. Researchers wishing to replicate the full extraction pipeline can collect equivalent review data under their own Google Maps API access in accordance with Google's Terms of Service.

Raw corpus data cannot be redistributed due to platform terms and privacy considerations. The released Gentags and derived experimental artifacts are sufficient to replicate the structural and decision analyses reported in this paper. Full pipeline replication from raw reviews requires an independently collected authorized corpus, for which all necessary prompts, scripts, and frozen definitions are provided.

11 References

Banko, Michele, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*.

Campos, Ricardo, Vitor Mangaravite, Arian Pasquali, Alipio M. Jorge, Celia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Chern, Shilong, Minghao Hu, Saurabh Mishra, Pratyay Banerjee, Mona Diab, Baolin Peng, Chunyuan Li, and Pengcheng He. 2023. FacTool: Factuality detection in generative AI—a tool-augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dong, Li, and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gao, Luyu, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hu, Xiangkun, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, and Yue Zhang. 2024. RefChecker: Reference-based fine-grained hallucination checker for large language models. *arXiv preprint arXiv:2405.14486*.

Lewis, Patrick, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Min, Sewon, Eric Zelikman, Yangsibo Huang, Yuntian Deng, and Luke Zettlemoyer. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Packer, Charles, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*, pages 1–20. John Wiley & Sons.

Salton, Gerard, and Christopher Buckley. 1988.

Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Wang, Wenya, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zettlemoyer, Luke S., and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Tengyu Ma. 2022. Least-to-most prompting enables complex reasoning. *arXiv preprint arXiv:2205.10625*.

A Implementation Details

This appendix documents the exact prompts, model configurations, decoding settings, validation rules, and frozen lexicons used throughout the experimental pipeline. All parameters were frozen before data collection and not modified during analysis.

A.1 Extraction Prompts

Three prompt variants are used for gentag extraction. All instruct the model to produce a JSON list of short semantic phrases grounded in the input reviews. They differ in how strongly they constrain grounding and brevity.

Minimal prompt.

> Extract semantic tags ("gentags") for this venue based on the reviews. A gentag is a short, meaningful semantic phrase (typically 1–4 words) that captures one idea expressed or strongly implied in the reviews. Include any gentags that describe atmosphere, food, service, vibe, crowd, or typical occasions mentioned in the reviews. Do not invent information beyond what the reviews support. Return only a JSON list of gentags.

Anti-hallucination prompt.

> Extract semantic tags ("gentags") for this venue based **ONLY** on what is explicitly stated or clearly implied in the reviews. A gentag is a short, meaningful semantic phrase (typically 1–4 words) that captures a single idea grounded in the review text. It must not be a full sentence. Do **NOT** infer, assume, generalize, or guess any information that is not directly supported by the reviews. If a concept is uncertain, ambiguous, or weakly implied, do **NOT** include it as a gentag. Include only gentags that reflect concrete statements in the reviews. Return only a JSON list of gentags.

Short-phrase prompt.

> Extract semantic tags ("gentags") for this venue that summarize the key ideas expressed in the reviews. A gentag must be a short phrase of 1–4 words that represents one clear semantic idea. Do not produce full sentences. Tags must be grounded in the content of the reviews and should not rely on assumptions or outside knowledge. Return only a JSON list of short gentags.

A.2 System Prompts

Table A1 summarizes the fixed system-level instructions used during extraction. Only OpenAI and Grok runs used explicit system prompts; Gemini and Claude used the user prompt alone.

A.3 Extractor Models

Decoding settings for extraction: temperature, `top_p`, and `max_tokens` are not overridden (provider defaults), except Claude which requires an explicit `max_tokens` (set to 8192).

A.4 Output Parsing and Validation

Extraction responses are parsed using three fallback strategies in order:

1. Direct `json.loads()` on the raw response.
2. Strip markdown code-fence wrappers (```` ```json ... ``` ````) and retry.
3. Extract the first balanced `[...]` bracket sequence using depth-first bracket tracking.

After parsing, tags exceeding four words are moved to a `tags_filtered_out` field. Empty and whitespace-only tags are stripped. If all three parsing strategies fail, extraction is marked `parse_error`. There is no automatic retry.

A.5 Judge Prompts

Tag-based systems (`gentag`, `rake`, `yake`, `tfidf`, `gentag_truncated`).

> You are a strict Decision Judge.
>
> Use ONLY the provided tags. Do NOT use external knowledge. Do NOT infer facts not present in the tags. Treat synonyms as NOT present unless they appear exactly in the tag list.
>
> Decision rules:
>
> - If the persona has a hard requirement and the requirement is VIOLATED based on the provided indicator set, output REJECT.
> - If the persona has a hard requirement and the requirement is SATISFIED, do NOT reject unless other clear blockers exist.
> - RECOMMEND if tags contain clear supports AND no clear blockers.
> - BORDERLINE if tags are mixed or ambiguous.
> - If the persona has no hard requirement, weigh all relevant tags.
>
> Return ONE line of valid JSON and nothing else, with fields:
>
> - `decision`: `REJECT`, `BORDERLINE`, or `RECOMMEND`
> - `requirement_status`: `SATISFIED`, `VIOLATED`, or `NOT_APPLICABLE`
> - `blockers`: `["..."]`
> - `supports`: `["..."]`
> - `tags_used`: `["..."]`
> - `justification`: one sentence
>
> Strict rules:
>
> - `tags_used` MUST be a subset of the provided tags (exact string match).
> - `blockers` and `supports` MUST be subsets of `tags_used`.
> - If a cited tag is not in the provided list, the response is INVALID.

Full-Evidence Reference (FER). The FER prompt follows the same decision rules but replaces "tags" with "reviews" and requires `evidence_quotes` (short quotes from reviews) instead of `tags_used`.

A.6 Judge Models and Decoding

Decoding settings: temperature and `top_p` use provider defaults. Max output tokens are hardcoded to 512 for judge responses.

A.7 Persona Indicator Lexicons

Hard requirements use frozen exact-match indicator sets. The judge checks whether any tag in the representation exactly matches an entry in the corresponding set.

Table A1: System prompts used by provider.

Provider	System Prompt
OpenAI, Grok	"You extract only JSON lists of gentags based on reviews. No explanations."
Gemini, Claude	None (user prompt only)

Table A2: Extractor models and pricing used in Phase 2.

Key	Model ID	Provider	Input ($/Mtok)	Output ($/Mtok)
openai	gpt-5-nano	OpenAI	0.05	0.40
gemini	gemini-2.5-flash	Google	0.25	0.50
claude	claude-sonnet-4-5	Anthropic	3.00	15.00
grok	grok-4	xAI	2.00	10.00

P1 Food Critic.
`negative_present_rejects`. **Positive:** delicious food, good food, excellent food, great food, quality food, fresh food, tasty food, amazing food, fresh ingredients, well-prepared food, flavorful food, outstanding cuisine, authentic flavor.

Negative: bad food, inconsistent food, poor food quality, cold food, undercooked, overcooked, raw and burnt, tasteless, bland food, stale food, low quality food, terrible food, disgusting food, flavorless.

P2 Sports Fan.
`indicator_present_not_reject`. watching game, watch games, watching sport, live sports, sports bar, sport bar, big screen, big screen sport, screen everywhere, game night, game-day vibe, sports viewing, favorite sport bar, large screen, TVs for sports, live game viewing.

P3 Quick Lunch Worker.
`indicator_present_not_reject`. fast service, quick service, quick bite, speedy service, rapid service, efficient service, fast food, prompt service, short wait, fast counter service, no wait, minimal wait, swift service, quick lunch.

P4 Balanced Diner. No hard requirement. Soft factors: food quality, service quality, ambiance.

A.8 Facet Anchors (Phase 3)

State-Gini analysis uses 10 frozen diagnostic facets. Tags are embedded with `text-embedding-3-large` (OpenAI, 3072 dimensions) and assigned to their closest facet anchor if cosine similarity $\geq \tau$ (default $\tau = 0.35$).

A.9 Aggregation and Scoring

Each Phase 5 condition (venue $\times$ persona $\times$ system) is evaluated with $N = 5$ repeated judge calls. Aggregation uses majority voting on the `decision` field. A minimum of three valid responses is required; conditions with fewer valid responses are marked `UNSCORABLE`. Ties are broken to `BORDERLINE`.

B Qualitative Audit of Gentag Failures

Out of 200 Gentag conditions (50 venues $\times$ 4 personas), Gentags disagree with FER in 41 cases.

The disagreement set supports a four-part failure-mode taxonomy:

- **FM1. Borderline Drift Under Mixed Evidence (33/41).** Gentags preserve both positive and negative propositions, and the judge falls back to `BORDERLINE` rather than committing. This is the dominant failure mode and reflects uncertainty induced by compressed gentag state under mixed evidence.
- **FM2. Exact-Match Indicator Misses (4/41).** Semantically relevant support exists (e.g., `fast delivery`, `game audio`) but the frozen exact-match indicator lexicon does not recognize it. This is partly a protocol artifact: the evaluation rule does not recognize semantically equivalent but lexically different tags.
- **FM3. Positive-Cue Anchoring (2/41).** An exact

Table A3: Judge models and pricing used in Phase 5.

Judge	Model ID	Input ($/Mtok)	Output ($/Mtok)
Primary	gpt-4o-2024-08-06	2.50	10.00
Cross-validation	claude-sonnet-4-20250514	3.00	15.00

Table A4: Frozen facet anchors used for Phase 3 assignment and State-Gini analysis.

Facet	Anchor phrase
food_quality	food quality, taste, freshness, delicious meals
coffee_drinks	coffee, espresso, latte, beverages, drinks
service	service quality, staff friendliness, speed, waiters
ambiance	atmosphere, ambiance, vibe, decor, cozy environment
price_value	price, value for money, affordable, expensive
crowding	crowded, busy, wait times, lines, availability
seating	seating, tables, outdoor patio, indoor space
dietary	dietary options, vegan, vegetarian, gluten-free
portions	portion size, generous servings, filling meals
location	location, parking, accessibility, neighborhood

positive indicator (e.g., `efficient service`) overrides contradictory negative evidence (`slow service`) in the same state. This suggests a concrete extension: conflict-aware state resolution.

- **FM4. Missed Negative Cue (2/41).** Negative semantics (`hygiene concern`, `unsanitary practice`) are present in the state but not surfaced in the decision. The representation preserves the relevant cue, but the downstream judge does not use it.

The audit shows that remaining errors are structured rather than arbitrary. Several exact reversals are protocol artifacts (frozen indicator lexicons) rather than representational failures. This sharpens the discussion: the dominant residual error is mixed-evidence drift, not hallucination or extraction failure.

C Supplementary Analyses

This appendix collects supplementary analyses, tables, and figures that support the main empirical trends reported in Section 5. Each subsection provides brief context for the corresponding supplementary material.

C.1 Stability (Phase 2)

Cross-prompt cosine similarity remains above 0.95 for all models, and cross-model cosine exceeds 0.94 for all prompt types. Together, these results support the main-paper claim that the recovered semantic state is not sensitive to the specific extraction configuration.

Per-model run-to-run bar charts show the same descriptive pattern as the main-text stability tables: extractor models differ in lexical repeatability, but all remain in the high-semantic-stability regime.

Retention measures how well the gentag state preserves the semantic content of the original reviews. All models achieve retention above the random baseline (+0.164), with Claude showing the highest median retention.

C.1.1 Evidence-induced Dispersion

Mean variability generally declines as more review evidence is available, falling from 0.0568 for venues with fewer than 200 review tokens to 0.0424 above 1000 tokens. The association is modest but statistically reliable (Pearson $r = -0.230$, $p = 0.00045$; Spearman $\rho = -0.263$, $p = 5.4 \times 10^{-5}$).

Table C1: Gentag variability by review-token bucket.

Token Bucket	Mean Variability	N Venues
< 200	0.0568	104
200–400	0.0465	87
400–600	0.0454	29
600–1000	0.0462	9
> 1000	0.0424	1

This pattern suggests that dispersion reflects iden-

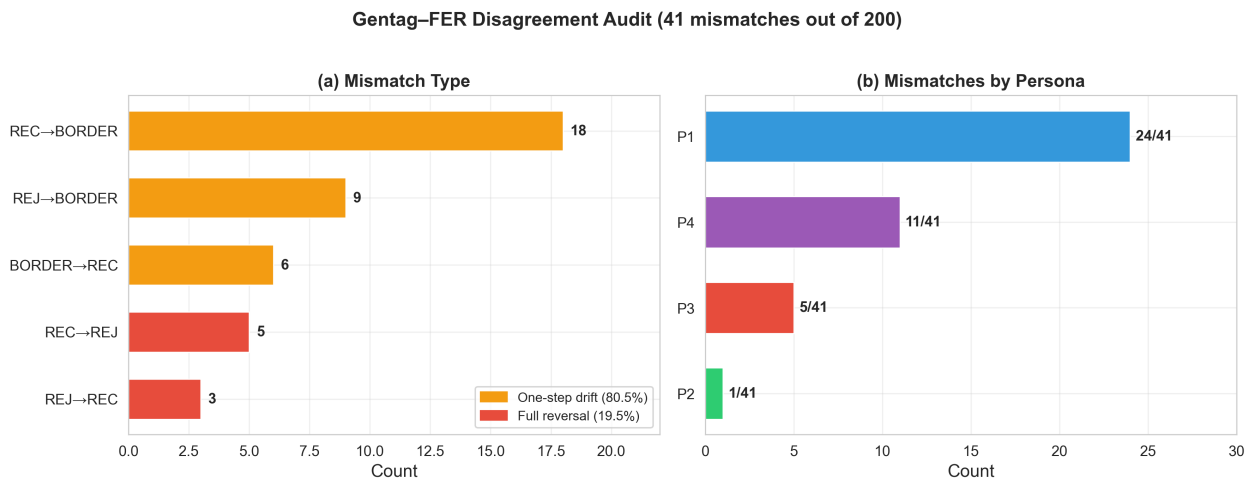

Figure B1: Gentag–FER disagreement audit. Left: mismatch types (80.5% one-step drift, 19.5% full reversals). Right: concentration by persona.

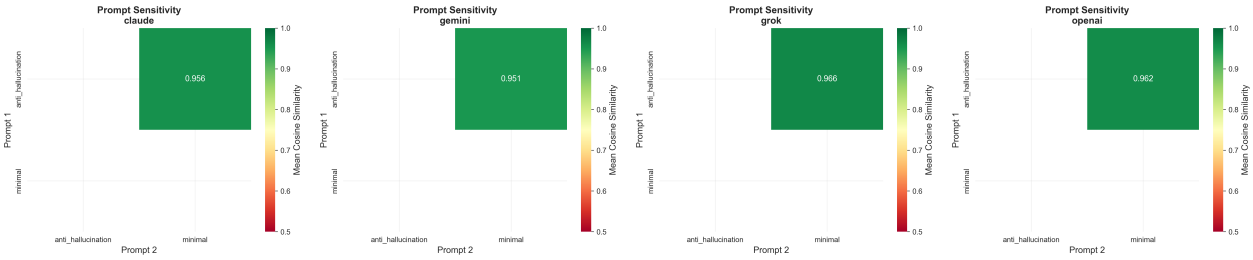

Figure C1: Cross-prompt semantic similarity heatmaps by model. All prompt pairs show cosine > 0.95.

tifiability rather than arbitrary noise: richer evidence makes the recovered semantic state easier to pin down across repeated runs.

C.2 Structure (Phase 3)

The structural advantage reported in Section 5.2 is robust to threshold choice. At all three tested thresholds ($\tau = \{0.30, 0.35, 0.40\}$), gentags retain more semantic mass within the facet inventory and show lower State-Gini than lexical baselines. The coverage gap widens at stricter thresholds, where baselines lose most of their assigned mass.

C.2.1 Bleed Check

Because the structural analysis uses hard argmax assignment, we also test whether the main pattern could be driven by noisy boundary cases. The bleed check measures the gap between the highest and second-highest facet similarities for each tag: larger gaps indicate cleaner assignments, while smaller gaps indicate tags closer to facet boundaries.

Table C2: Gentag bleed-check summary under facet assignment.

Gentag metric	Value
Mean primary–secondary gap	0.065
Median gap	0.039
Near-miss rate (gap < 0.05)	57.4%
Clear-primary rate (gap ≥ 0.10)	20.5%
Mean primary similarity	0.343

For Gentags, facet boundaries are often soft rather than sharply separated: a majority of tags show small top-two margins, and only about one fifth have a clearly dominant facet under the current criterion.

This comparison is informative for two reasons.

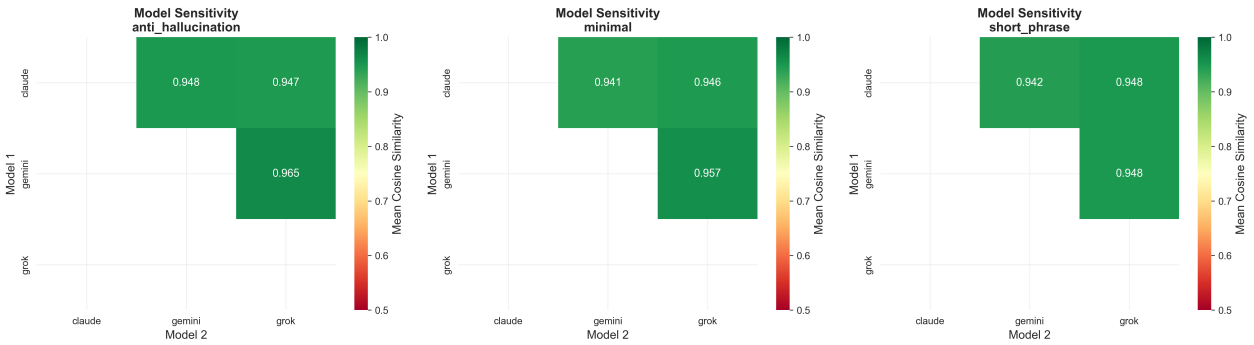

Figure C2: Cross-model semantic similarity heatmaps by prompt. All model pairs show cosine > 0.94.

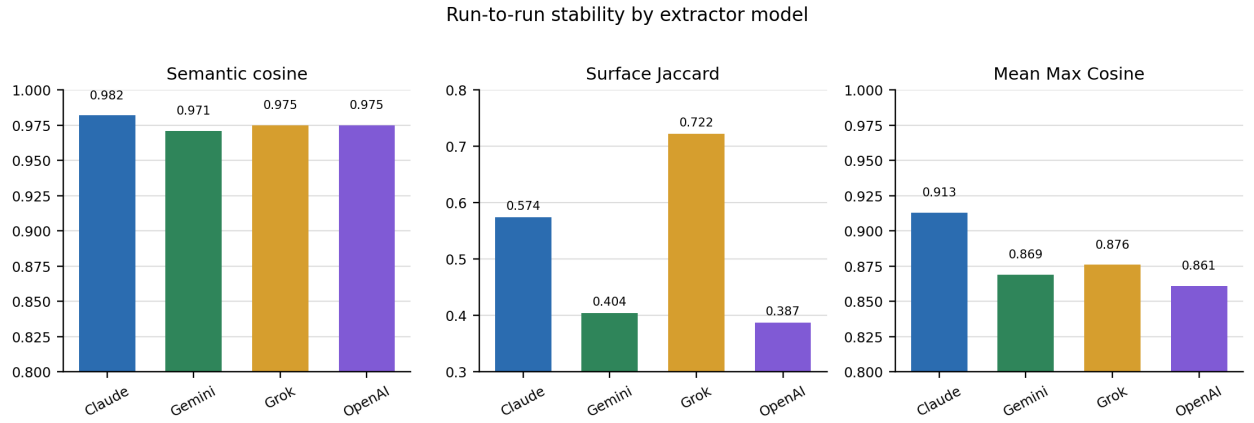

Figure C3: Per-model run-to-run stability comparison for gentag extraction metrics.

Table C3: Bleed-check comparison across gentags and lexical baselines.

Method	Mean gap	Median gap	gap < 0.05	gap ≥ 0.10	Mean primary sim
Gentags	0.065	0.039	57.4%	20.5%	0.343
RAKE	0.056	0.030	64.0%	18.0%	0.319
YAKE	0.056	0.030	64.7%	18.1%	0.318
TF-IDF	0.055	0.025	66.9%	17.4%	0.321

First, Gentags are not uniquely ambiguous; lexical baselines are slightly more boundary-ambiguous by this diagnostic. Second, it weakens a simple objection to the structural result: lower Gentag State-Gini is not explained by unusually noisy argmax assignments. If anything, baselines have smaller top-two margins and higher near-miss rates.

These diagnostics do not make the facet inventory a clean ontology; all methods show soft boundaries. They do, however, clarify how to read the structural analysis: facets function as a diagnostic probe space, and Gentags perform at least as well as baselines on ambiguity while achieving substantially better facet coverage. This motivates joint reporting of `other_rate`, State-Gini, and gap-based ambiguity.

Across the 10 diagnostic facets, Gentags spread mass across `food_quality`, `service`, `ambiance`, and other facets, while lexical baselines concentrate in the "other" bucket (67–69% of mass unassigned). This directly visualizes the facet-coverage advantage discussed in Section 5.2.1.

C.3 Decision (Phase 5)

C.3.1 Decision Entropy

Decision entropy provides a complementary view of representational quality. A useful representation should not only improve correctness, but also produce coherent and decisive downstream behavior.

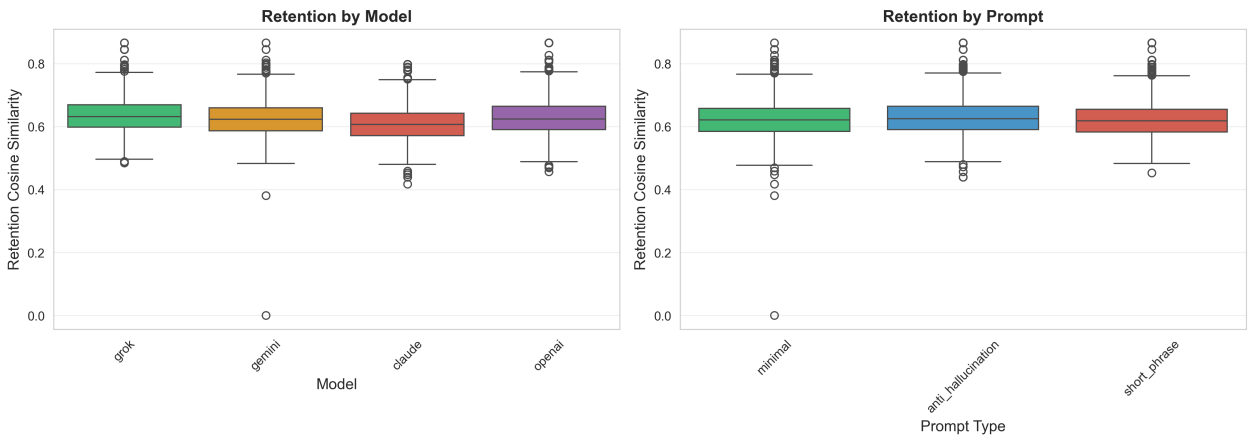

Figure C4: Source retention by model and prompt.

Table C4: Decision entropy and distributional divergence from FER by system.

System	H	P(REJ)	P(BORD)	P(REC)	L_1 vs. FER
FER	1.393	52.5%	12.5%	35.0%	—
Gentag	1.506	49.0%	23.0%	28.0%	0.210
Gentag truncated	1.520	47.7%	24.6%	27.6%	0.242
RAKE	1.500	49.5%	28.3%	22.2%	0.316
YAKE	1.460	49.5%	34.0%	16.5%	0.430
TF-IDF	1.430	49.8%	36.2%	14.1%	0.474

Lexical baselines systematically shift probability mass away from `RECOMMEND` and toward `BORDERLINE`. This is consistent with semantically incomplete or opaque representations: they do not only increase error, they also induce uncertainty. Gentags remain substantially closer to the full-evidence FER decision distribution.

Decision distributions show the same pattern as the entropy table: lexical baselines overproduce `BORDERLINE`, while Gentags remain closer to the FER reference distribution.

Cross-judge agreement (Cohen's kappa) remains substantial across systems. The overall kappa of 0.712 indicates substantial agreement between the two judge models. `gentag`, `FER`, and `gentag_truncated` show the most stable cross-judge agreement, while TF-IDF and YAKE show slightly lower kappa, consistent with fragment-based representations being harder to interpret consistently across different judge models.

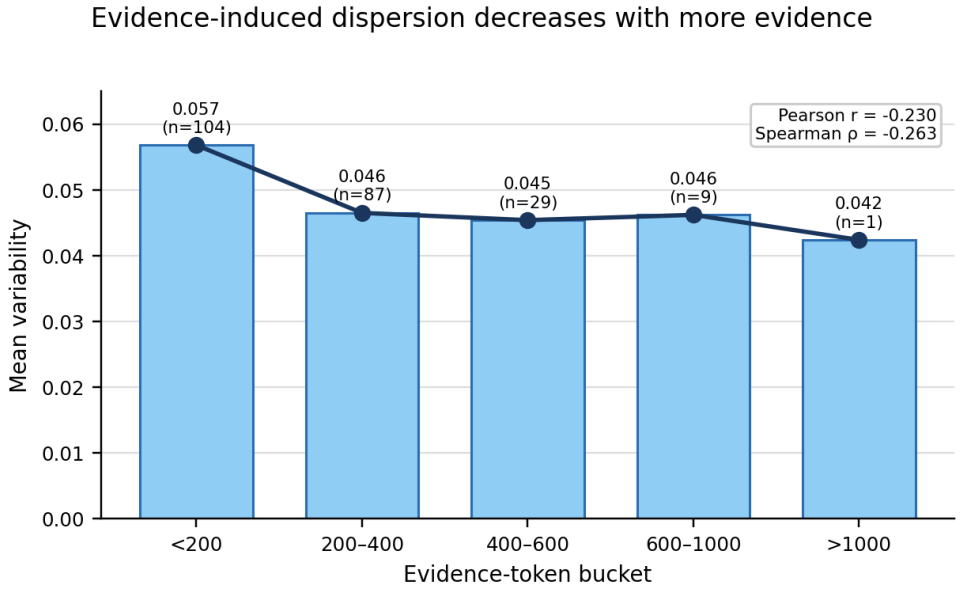

Figure C5: Evidence-induced dispersion by review-token bucket: variability decreases as evidence availability increases.

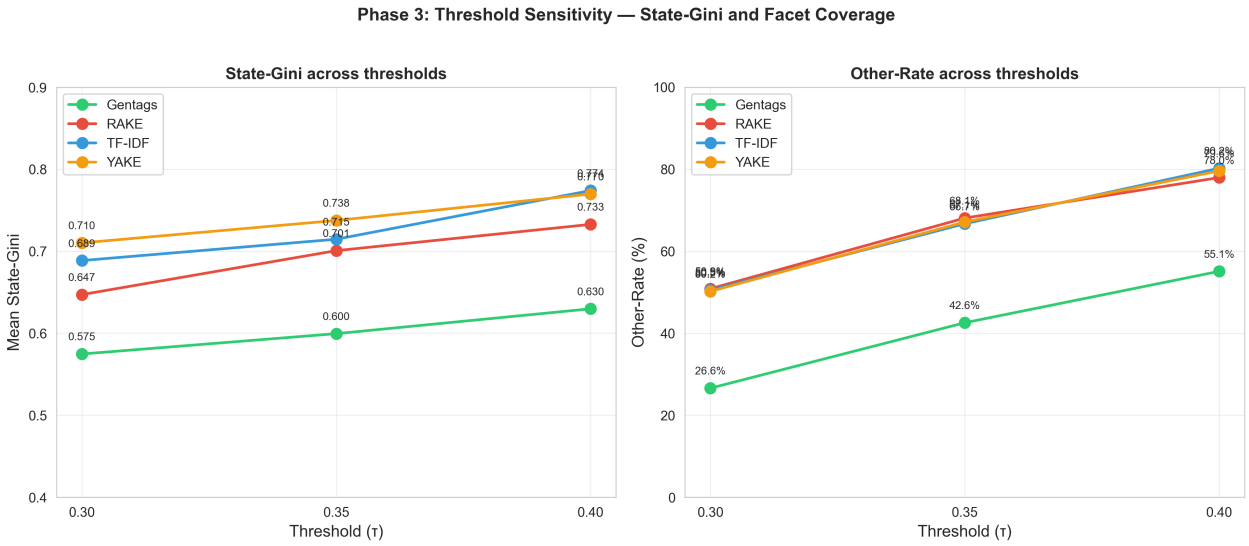

Figure C6: Threshold sensitivity—State-Gini and other-rate across $\tau = \{0.30, 0.35, 0.40\}$.

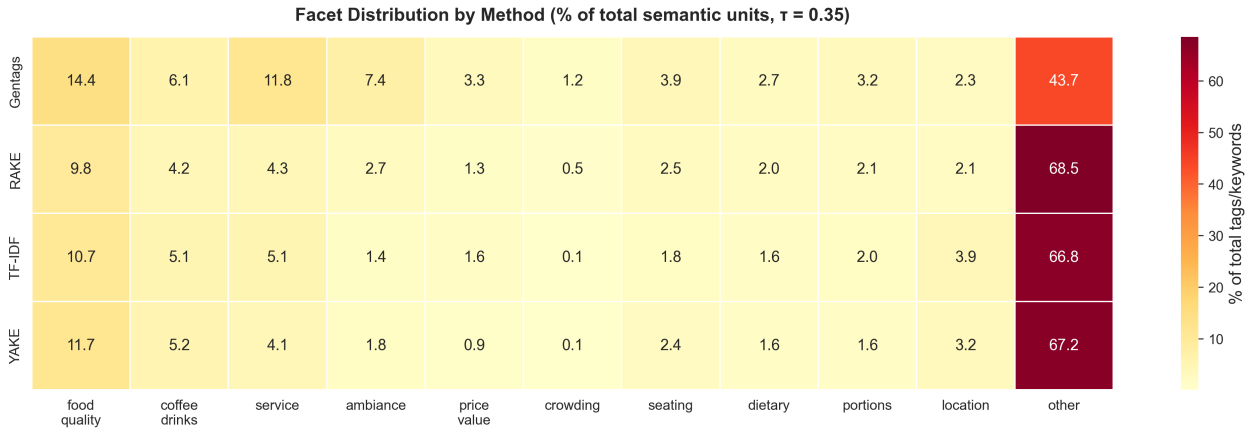

Figure C7: Per-facet distribution by method (% of total semantic units, $\tau = 0.35$).

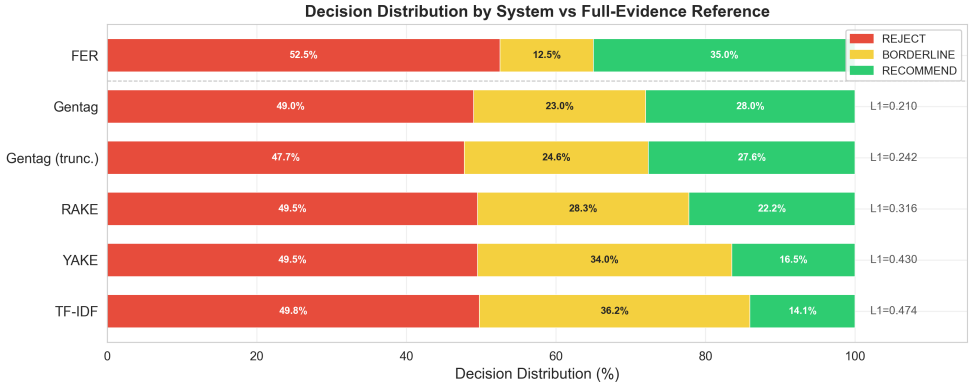

Figure C8: Decision distribution comparison showing BORDERLINE bloat in lexical baselines and tighter Gentag alignment to FER.

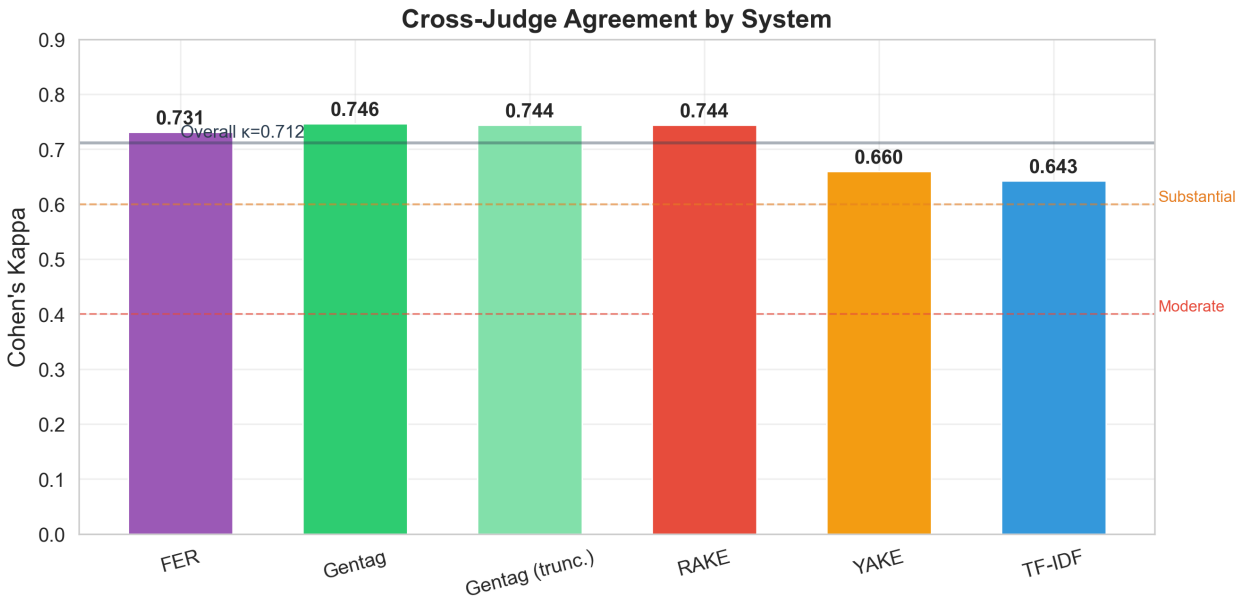

Figure C9: Cross-judge Cohen's kappa by system. All systems show substantial agreement ($\kappa > 0.6$). Overall $\kappa = 0.712$.