The Validity–Fidelity Gap: A Controlled Study of Representation Drift in Structured Extraction Under Redundant Contexts

Gokul Srinath Seetha Ram
Independent Researcher
s.gokulsrinath@gmail.com

Anirudh Vasudevan
Independent Researcher
anirudhvasudevan11@gmail.com

Rashmi Elavazhagan
Independent Researcher
rashmie30@gmail.com

Abstract

Pipelines that use LLMs for structured extraction are commonly monitored by checking that each output parses as JSON and matches the expected keys. We show that this check can stay perfect while the extracted content degrades. We call the failure **representation drift**: the model still finds an entity but writes it in a different form than the intended one, for example a nested key–value object where a plain string is expected. The output is still valid JSON, yet the entity is scored both as missing (a false negative) and as an unexpected extra (a false positive). We study this with synthetic clinical notes repeated verbatim 1×–32×, which gives exact ground truth with prompt, schema and decoding held fixed. For LLaMA-3.1-8B-Instruct, micro-F1 falls from 0.499 to 0.259 while JSON validity stays 1.00. All 280 medication entities scored as missing in a 100-patient run are present in the output as nested objects, and a follow-up request with no schema recovers 96.7% of medications versus 40.0% under the schema at 16×. We introduce the Differential Redundancy Effect (DRE), the extra F1 drop under repeated content relative to an approximately length-matched unrelated-filler control, and use it as a diagnostic contrast rather than a causal estimate. The effect is scale-dependent: DRE is large at 8B and near zero at 70B and Qwen2.5-72B, where F1 does not degrade. Validity metrics stay at 1.00 in both regimes, so they carry no information about content fidelity. A smaller synthetic news experiment shows the same dissociation with a different schema. We also give a zero-ground-truth drift check and monitoring recommendations. All evidence is synthetic, and we make no claim about real clinical notes.

1 Introduction

Structured extraction pipelines map natural language documents to typed schemas — lists of conditions, medications, procedures, observations. These pipelines are increasingly built on instruction-tuned LLMs (Brown et al., 2020; Ouyang et al., 2022). A central question in structured-generation benchmarks is whether outputs are *structurally valid*: does the response parse as JSON, and does it contain the required keys and array types? (Geng et al., 2025; Tang et al., 2023; Yang et al., 2025) Other work asks how format restrictions affect task quality (Tam et al., 2024). In production, structural validity rates are also a natural reliability signal, logged in dashboards and used to gate downstream processing; we ask whether that signal is informative about content fidelity.

Our question is: *can structural validity be perfect while content fidelity silently fails?* In the settings we tested, yes, and we characterize how.

The phenomenon: representation drift. Under redundant long contexts, an instruction-tuned model can keep *what* it was asked to extract while changing *how* it writes it. Table 1 shows a real case: the schema example in the prompt has empty arrays for each field, and the reference answers are plain strings, but the model writes each medication as a JSON object. A JSON or schema validator sees a valid array; a downstream consumer expecting entity strings gets records. We call this **representation drift**. We use "format contract" below to mean this expected representation (plain strings in array fields); it is implied by the schema example and our gold data, not declared as a type in the prompt (see Limitations).

Under this view, drift has a characteristic diagnostic signature: FP and FN counts inflate together at the entity-string level, while structural metrics (JSON validity, schema shape, and the Structural Drift Score, SDS, defined in Section 3) stay at their ideal values. It is therefore invisible to any monitoring that uses only structural features. At 70B scale we do not see F1 degradation; larger models can improve extraction recall with more context. Structural monitoring is equally uninformative there:

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2112–2126
November 6-10, 2026 ©2026 Association for Computational Linguistics

Gold	`"Albuterol inhaler 90 mcg 2 puffs q4-6h PRN"`
Model output (8B, 16×)	`{"name": "Albuterol inhaler", "dosage": "90 mcg", "frequency": "2 puffs inhaled q4-6h", "reason": "PRN wheeze"}`
Validators	JSON valid: yes; schema valid: yes (SDS = 0)
Content scored	gold string missing (FN) and object unmatched (FP)

Table 1: A representation-drift case (patient_01, run_01; the full output array is in the released artifacts). The information is present; only its representation differs, so all structural checks pass.

whether F1 goes up, down, or stays flat, validity metrics hold at 1.00.

We read these results as a hypothesis about instruction following under load, not an established mechanism: if format-level failures precede recall failures, they should depend on accumulated content more than string identity and hit high-cardinality fields first. Our 8B experiments are consistent with this, but they are behavioral and test no mechanism.

Study design. We use synthetic clinical notes with verbatim repetition (following Liu et al. 2024's controlled-synthetic methodology): exact repetition holds all variables fixed except the stressor and makes ground-truth FN/FP measurement unambiguous. Redundancy is the controlled stressor that exposes the validity–fidelity gap, which is the object of study: structural and semantic metrics are computed on every call. It is a diagnostic stressor, not a model of real documents (see Limitations).

Main findings. We separate a claim that holds in every setting we tested from claims that depend on model scale.

- **Robust: structural validity is uninformative about fidelity.** Across three models (8B, 70B, Qwen2.5-72B), all repetition levels, and two extraction domains (clinical notes, news articles), JSON validity is 1.00 and SDS is 0, whether content F1 degrades (8B), improves (70B), or stays flat.
- **At 8B: drift is a change of representation, not lost entities.** (i) All 280 FN-scored medications (100-patient run; 5× and 16×) are present in the model output as nested-object entries. (ii) A free-form medication-listing probe with no schema recovers 96.7% recall from the same model that reaches 40.0% under structured extraction at 16× (+0.567 gap). The entity information remains accessible; what changes is its representation.
- **At 8B: degradation is threshold-like and depends on accumulated content.** The 2×→5× step accounts for 55% of the F1 loss. Section-level redundancy alone does not reproduce the threshold at 5×. Shuffled and paraphrased redundancy do not remove the effect (DRE 0.305 at 10× for paraphrase vs. 0.204 verbatim; the paraphrase set has a mention-loss caveat). We call the hypothesized driver "semantic saturation"; it is a descriptive label, not a measured mechanism.
- **Scale attenuates the effect.** At 8B, F1 *degrades* with redundancy and DRE is large. At 70B and Qwen2.5-72B, DRE is near zero and F1 improves under both redundancy and filler. We therefore do not claim that redundancy causes drift in general; the claim is restricted to the smaller model.
- **A zero-ground-truth check exists.** A character-scan heuristic flags likely drift without labels. It is a trip-wire to prompt a labeled evaluation, not a severity metric.

2 Related Work

Controlled long-context probing. A growing body of work uses controlled synthetic benchmarks to expose failure modes that real-world corpora obscure. Liu et al. (2024) construct multi-document QA tasks where the relevant passage is placed at varying positions, demonstrating that LLMs disproportionately attend to context boundaries and miss information buried in the middle. We adopt the same controlled-synthetic methodology and apply it to redundancy rather than position: our stressor is *accumulating semantic content* rather than *positional distance*. Crucially, in our setting retrieval does not fail — the model identifies the correct entities — yet output format breaks down, revealing a dissociation that positional probes do not surface. Needle-in-a-haystack tests (Kamradt, 2023) stress retrieval under distraction; ours stresses representation under content saturation, a complementary axis. Bai et al. (2023), Hsieh et al. (2024), and Tay et al. (2021) provide multi-task long-context evaluation suites that measure QA, summarization,

and retrieval accuracy; to our knowledge they do not report structured-output format fidelity separately from content correctness. Architectural and positional-encoding advances have extended nominal context windows (Beltagy et al., 2020; Su et al., 2024; Press et al., 2022; Peng et al., 2023); our 8B results suggest that a model processing a 32× note still extracts the content but represents it differently, which we read as an instruction-following effect rather than a retrieval failure.

Structured generation and validity vs. content. Benchmarks for structured generation (Geng et al., 2025; Tang et al., 2023; Yang et al., 2025) measure, among other things, whether outputs are syntactically valid and schema-conformant (for example, Geng et al. 2025 compare constrained-decoding engines on compliance, coverage and quality over real-world schemas, with the input held fixed). Complementarily, Tam et al. (2024) study the converse direction: how *imposing* format restrictions affects downstream task performance, finding that stricter format constraints can degrade reasoning. Our question is different from both: holding the format requirement fixed, we ask whether structural validity *metrics* remain informative about content fidelity, and find that validity can remain at 1.00 while content fidelity changes with context length, model scale, and field type. Constrained decoding (1rgs, 2023; Willard and Louf, 2023; guidance-ai, 2023) enforces syntactic validity by construction, which might suggest it eliminates the problem. We argue conceptually (we did not test constrained decoders) that it may not: a decoder given a schema that leaves array items unspecified admits objects, and one that fixes items to strings still admits a string whose *content* is a serialized record. In such pipelines, the most natural monitoring signal (does the output parse?) is orthogonal to the failure we describe.

Instruction following under load. Instruction following — obeying stated constraints on format, length, style, and content — is a fundamental capability of instruction-tuned models (Ouyang et al., 2022; Wei et al., 2022a,b; Zhou et al., 2023). Work on complex instruction following has shown that models can fail to satisfy all stated constraints simultaneously, particularly when constraints are numerous or in tension. Shi et al. (2023) show that *irrelevant* context disrupts arithmetic reasoning; we extend this observation to the structurally different setting where *highly relevant* semantic content accumulation disrupts format instructions while leaving content knowledge intact. This distinction matters: Shi et al. (2023)'s result can be explained by retrieval confusion (distractor text competes with relevant content), whereas representation drift occurs even when the model demonstrably knows what to extract. Work on attention sinks (Xiao et al., 2024) provides a candidate mechanism — instruction tokens lose relative attention weight as content tokens fill the window — but we treat this as a hypothesis rather than a measured account, since mechanistic attribution would require intervention experiments beyond our scope.

Hallucination, faithfulness, and calibration. Representation drift occupies a distinct position in the taxonomy of LLM output failures. Factual hallucination (Dziri et al., 2022; Ji et al., 2023; Lin et al., 2022) involves the fabrication of content not supported by the input; representation drift involves the *correct identification* of content followed by its encoding in a non-compliant format. Self-consistency methods (Manakul et al., 2023) detect hallucination by checking agreement across samples; they cannot detect drift because repeated samples will *consistently* produce the same nested-dict format under high redundancy — agreement is high, but format compliance is broken. Our dissociation probe gives direct behavioral evidence that, at 8B, the failure is in the representation rather than in the information the model has access to.

Clinical and general information extraction. LLMs have been applied to clinical IE and NER tasks (Agrawal et al., 2022; Gilardi et al., 2023; Savova et al., 2010; Wang et al., 2023), often evaluated on whether the correct entities are present in the output with little attention to output schema compliance. Our work shows that entity presence (as measured by recall under lenient matching) can coexist with schema non-compliance in ways that break downstream consumers without triggering any standard evaluation alarm. Retrieval-augmented generation (RAG) pipelines (Lewis et al., 2020) introduce a particularly natural setting for representation drift: when multiple retrieved passages describe the same entity, the effective redundancy of that entity in the context window increases with retrieval depth. RAG-based extraction may therefore be exposed to a similar effect even when no individual document is long; we did not test this.

3 Setup

Data. 10 synthetic one-page clinical discharge summaries, written as fixed templates in the style of Synthea records (Walonoski et al., 2018) (they are not output of the Synthea generator), with ground truth for the schema `{patient_id, conditions[], medications[], observations[], procedures[]}`. The "100-patient" sets used for large-N results cycle these 10 templates, so they contain 10 distinct documents, not 100 (see Section 6 and Limitations). Repetition factors are $r \in \{1, 2, 5, 10, 16, 32\}$ (the base note concatenated r times). The filler control keeps the note once and appends randomly chosen non-clinical filler paragraphs (four fixed sentences) until a character-based length estimate matches the redundancy condition. **Models.** LLaMA-3.1-8B/70B-Instruct (Grattafiori et al., 2024) and Qwen2.5-72B-Instruct (Yang et al., 2024) through the DeepInfra API; temperature 0, 3 runs/condition for the 8B pilot (1 run for most other conditions; Appendix Table 14).

Three levels of validity. We distinguish three levels at which an output can be checked. (1) **JSON validity**: the raw response parses as a JSON object (after extracting the outermost object and removing any markdown fence). (2) **Schema validity**: the required keys are present, no extra keys appear, and the array fields are arrays. SDS and the repair rate below measure levels (1) and (2). (3) **Content-level type fidelity**: every array element is a string naming one entity in the intended surface form. Nothing in levels (1) and (2) checks this, and a JSON object inside an array passes both; representation drift is a failure at level (3).

Metrics. *Entity matching and semantic metrics.* Predicted and gold entities are compared as sets within each of the four array fields. Each element is lower-cased, whitespace is collapsed, and punctuation other than letters, digits, spaces and `-/().%` is removed; non-string elements (e.g. objects) are converted with Python's `str()`, so they never equal a gold string. A predicted string matching a gold string is a TP; an unmatched predicted string is an FP; an unmatched gold string is an FN. Micro precision/recall/F1 are computed over the four fields pooled. FN counts include entities that appear in a non-string form, which is why drift produces paired FN and FP. *Structural.* (i) **JSON validity**: per-call binary, level (1) above; (ii) **repair rate**: fraction of calls needing normalization before schema comparison; (iii) **Structural Drift Score (SDS)**: a per-call weighted penalty, $\text{SDS} = 2.0\,[\text{malformed JSON}] + 0.5\,\#\text{missing keys} + 0.25\,\#\text{extra keys} + 0.5\,[\text{schema mismatch}] + 0.25\,[\text{markdown fence}]$, so $\text{SDS} = 0$ iff the output is fully conformant at levels (1) and (2). Full design, prompts, evaluator variants, and the filler audit are in Appendix A.

Rep	Entity type	TP	FP	FN	Prec.
1×	conditions	13	5	6	0.72
1×	medications	2	23	23	0.08
1×	observations	15	5	7	0.75
1×	procedures	3	4	4	0.43
32×	conditions	12	5	7	0.71
32×	medications	0	25	25	0.00
32×	observations	4	17	18	0.19
32×	procedures	3	4	4	0.43

Table 2: Entity-type breakdown at 1× vs. 32× (run_01, 8B pilot). Symmetric FP/FN inflation in medications (0 TP; 25/25 FP/FN at 32×) is the signature of representation drift. Medications are already affected at 1× (2 TP; 23 FP/FN). Conditions and procedures are stable across repetitions. Novel insertions (FP without matching FN) are rare.

4 Characterizing Representation Drift

Extraction errors decompose into three types:

1. **Genuine omission**: entity absent from output entirely (FN, no matching FP).
2. **Representation drift**: entity present in output as a nested object instead of the required plain string — produces a paired FN + FP simultaneously.
3. **Novel insertion**: hallucinated FP with no corresponding gold entity.

Drift produces the signature of **symmetric FP/FN inflation**: when `"acetaminophen 650 mg q6h prn"` is emitted as the object `{"name":"acetaminophen", "dose":"650mg",...}`, one FN and one FP appear together.

Entity-type breakdown. Table 2 shows the pattern. **Medications** collapse at 32× (TP: 2→0; FP/FN both reach 25): the entities are still emitted, but not as plain strings. **Conditions** and **procedures** are invariant; **observations** degrade partially. At 32×, medication arrays contain objects (`{"name":"acetaminophen",`

Condition	Mean recall
Structured extraction at 1×	0.833
Structured extraction at 16×	0.400
Free-form probe (1× document)	**0.967**
Gap: free-form − structured 16×	+0.567

Table 3: Dissociation probe results (30 patients; 8B model). When asked to list medications in plain text without a schema, the model recalls 96.7% of gold medications — substantially above both its 1× structured baseline (83.3%) and its 16× structured performance (40.0%). The +0.567 gap quantifies the representation shortfall: the entities are accessible via a schema-free probe (on the 1× document), but are not written as plain strings under the schema at 16×.

`"dose":"650mg",...}`) that a string-expecting consumer cannot use directly, although the output is structurally valid. (The pilot shows some object-valued medications already at 1×, so drift amplifies a pre-existing tendency.)

Direct dissociation evidence. **Part 1** (zero-cost analysis of existing runs): across 280 FN cases in the 100-patient run (130 at 5×, 150 at 16×), **100% of FN-scored medications are present in output as nested-object entries**, i.e. none of these FNs is a genuine omission. **Part 2** (30 free-form follow-up calls): prompt "list medications, no JSON" sent to the same model on the same 1× documents. Results (Table 3):

Free-form recall (0.967) exceeds even the 1× structured baseline (0.833). The 16× shortfall (0.400) is therefore not explained by the model lacking the entities: they are recoverable without the schema. Note that the probe uses the 1× document and no schema, so it shows that the information is accessible, not that the 16× context would have been answered correctly in free form. **Both analyses point the same way: at 8B, drift is a change of representation, not a loss of the entities.**

5 Redundancy Scaling Results (8B)

Table 4 and Figure 2 show the aggregate phenomenon; Figure 3 shows the redundancy vs. filler comparison. The simultaneous conditions — F1 falling, FN and FP rising symmetrically, JSON validity fixed at 1.00, SDS at 0 — constitute the operational definition of the validity–fidelity gap in this setting. A content-level check does track the decline: the brace/object scan of Section ef-sec:implications flags 4/30 calls at $1imes$ and 24/30 at $32imes$.

Rep	F1 ↑	FN ↓	FP ↓	JSON valid
1×	0.499	3.83	3.53	1.00
2×	0.457	4.10	3.70	1.00
5×	0.325	5.00	4.60	1.00
10×	0.311	5.07	4.77	1.00
16×	0.292	5.20	4.90	1.00
32×	0.259	5.40	5.10	1.00

Table 4: Redundancy scaling (8B; mean over 10 patients × 3 runs). F1 degrades monotonically; FN and FP inflate symmetrically (representation drift signature); JSON validity is perfect throughout. SDS = 0 at all levels (not shown).

A threshold, not a slope. The 2×→5× transition accounts for **55% of all F1 degradation** (ΔF1 = −0.132); subsequent levels contribute only 0.01–0.03 each. Beyond 5×, the model stabilizes in a persistent drift state. The threshold holds per-patient: 5 of 8 affected patients show their largest drop at or before 5×; 2 patients show zero degradation at all levels (Appendix C, Tables 15, 16).

6 Differential Redundancy Effect

To separate redundancy-specific effects from generic long-context pressure, we define the **Differential Redundancy Effect**:

$$\mathrm{DRE}(r) = \Delta_{\mathrm{red}}(r) - \Delta_{\mathrm{fill}}(r),$$

where $\Delta_c(r) = \mathrm{F1}(1, c) - \mathrm{F1}(r, c)$ is the within-patient F1 drop. Positive DRE means repeated evidence is associated with a larger F1 drop than an approximately length-matched unrelated filler. Computed separately for each model, this is what we call **cross-model DRE** (Table 9). Because the filler is sized by a character-count heuristic rather than matched on true prompt tokens (Appendix A, Table 13), we interpret DRE throughout as a *diagnostic contrast* against the implemented baseline, not a causal estimate under perfectly matched token length.

Table 5 shows the pilot DRE values; positive midrange DRE is replicated in the 100-patient large-N setting (Table 6; bootstrap 95% CI (Efron and Tibshirani, 1993) excludes zero at all midrange levels; Figures 7 and 8 in Appendix), so the effect is not specific to a single run. The absolute F1 drop under redundancy behaves similarly (Table 20; Appendix, Figure 6): at 16×, the mean paired F1 drop is 0.190 with a patient-level CI of [0.154, 0.227]. *Caveat on independence.* The 100-patient sets repeat each of the 10 templates ten

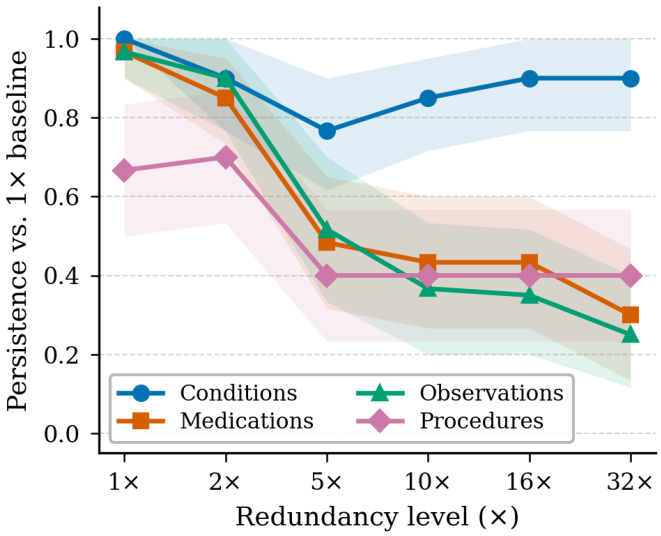


Figure 1: **Entity-type persistence relative to the 1× baseline, with 95% bootstrap CI bands (8B pilot; 10 patients).** Each line is the fraction of gold entities of that type still extracted as a plain-string TP. **Medications** (orange) and **observations** (green) drop steeply at 2×→5× (to ≈0.48 and 0.51) and reach 0.30 and 0.25 by 32×. **Conditions** (blue) are the most stable (0.77 at 5×, 0.90 at 32×). **Procedures** (pink) start near 0.67 and settle at 0.40 from 5×. Degradation is selective: the affected fields are those with richer sub-structure (dose/route/frequency; value/unit).

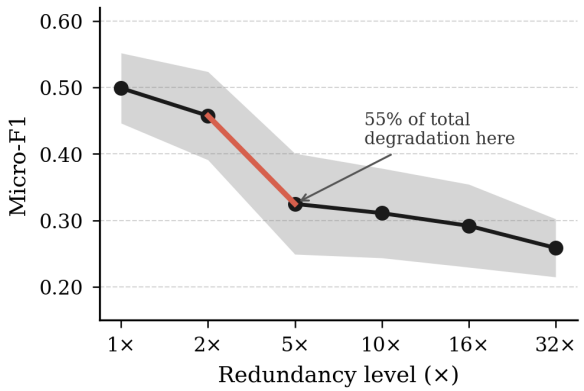


Figure 2: Micro-F1 vs. repetition factor (8B pilot). Monotonic degradation while structural validity (not shown) stays at 1.00: the validity–fidelity gap in aggregate.

Rep	$\Delta F1$ (red.)	$\Delta F1$ (filler)	DRE
1×	0.000	0.000	0.000
2×	0.042	−0.029	+0.070
5×	0.174	−0.016	+0.190
10×	0.188	−0.016	+0.204
16×	0.207	+0.012	+0.195
32×	0.240	0.240	+0.001

Table 5: DRE in the 8B pilot. Positive midrange DRE (5×–16×) indicates that exact repeated evidence is associated with more extraction degradation than approximately length-matched unrelated filler. Convergence at 32× reflects a shared long-context floor.

times, so patient-level bootstrap CIs treat repeated documents as independent and are too narrow. Re-bootstrapping over the 10 templates gives wider percentile intervals that still exclude zero for full-document DRE (5× [0.07, 0.32]; 10× [0.07, 0.30]; 16× [0.06, 0.29]) and for the F1 drop (16× [0.08, 0.31]). Effective sample size is 10 documents.

7 Semantic Content or String Identity?

Three analyses are consistent with the stressor being accumulated *semantic content*, more than exact string form, at 8B; none isolates a mechanism.

Shuffled redundancy. Shuffled copies (same content, permuted paragraphs) still produce F1 ≈0.08–0.22 below filler (Appendix, Table 17); breaking exact character identity yields only modest improvement.

Section-level redundancy. Repeating only the highest-density sections (diagnoses, medications) while the rest appears once (100 patients) fails to trigger the threshold: at 5×, section-level DRE = 0.013 (CI: [−0.014, 0.038]; CI crosses zero), while full-document DRE = 0.186 (Table 7; Appendix, Figure 5). Local entity-section repetition alone did not reproduce the 5× effect. At 10×–16× section-level DRE is small but positive at the patient level; with the template-level bootstrap its intervals include zero (10× [−0.02, 0.18]; 16×

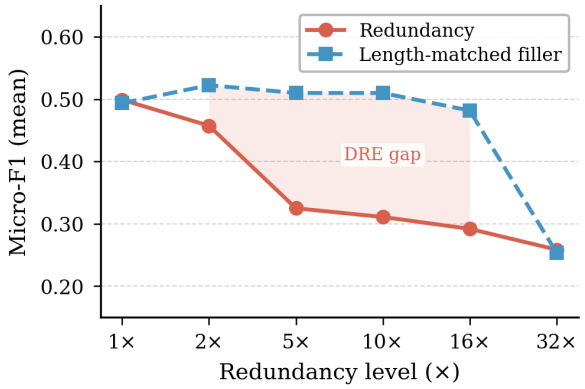


Figure 3: Redundancy vs. filler F1 (8B; filler only approximately length-matched). The gap between curves at midrange repetitions is DRE; convergence at 32× is the shared degradation floor.

Rep	Mean DRE	CI_{95}	n_+	n_-	n_0
5×	0.186	[0.144, 0.230]	48	0	52
10×	0.174	[0.133, 0.217]	53	6	41
16×	0.174	[0.132, 0.216]	55	9	36

Table 6: Large-N (100 patients, 8B) DRE with 95% bootstrap CI (paired by patient). $n_+/n_-/n_0$: patients with positive, negative, and zero DRE. Patient-level CIs exclude zero at all midrange levels (template-level CIs are in the text); no patient has negative DRE at 5×.

[−0.003, 0.17]), whereas full-document DRE does not.

Paraphrase and prompt robustness. Paraphrased redundancy (50 patients) does not attenuate DRE: 10× DRE = 0.305 vs. 0.204 verbatim (Appendix, Table 18; caveat: 48% mention loss). DRE is positive at 5× under three distinct prompt wordings (Appendix, Table 19). Together these results are consistent with a role for cumulative semantic content, but they do not establish it as the cause.

8 The Phenomenon at Scale

What scale tells us. Table 9 shows that cross-model DRE is near zero at 70B/72B, so the redundancy-specific effect we see at 8B attenuates with scale. Table 8 shows why: at 70B, F1 *improves* under both conditions (redundancy: 0.463→0.551; filler: 0.467→0.536) — larger models appear to benefit from additional context rather than showing the 8B pattern. DRE ≈ 0 because *both* conditions improve similarly, not because both degrade.

The invariant across all scales: JSON validity = 1.00, SDS = 0 throughout — structural monitoring is blind whether F1 is going down (8B) or up (70B).

Condition	5×	10×	16×
Full-document	0.186*	0.174*	0.174*
Section-level	0.013	0.073*	0.078*

Table 7: Full-document vs. section-level DRE (8B; 100 patients; *CI excludes 0). Section-level does not reproduce the effect at 5×. Stars are patient-level CIs; template-level CIs for section-level DRE include zero at all three levels.

Rep	F1 (redundancy)	F1 (filler)
1×	0.463	0.467
5×	0.551	0.536
10×	0.550	0.552
16×	0.550	0.558

Table 8: 70B 100-patient F1 by repetition (both conditions). Unlike 8B, F1 *improves* from 1× to 5× under both conditions and then stabilizes. DRE ≈ 0 because both conditions benefit similarly from additional context. JSON validity = 1.00 and SDS = 0 throughout (not shown) — structural metrics are equally uninformative whether F1 is going up or down.

We use "semantic saturation" only as a descriptive hypothesis for the 8B pattern, consistent with, but not tested against, instruction-token attention dilution (Xiao et al., 2024; Shi et al., 2023); we made no attention or internal-state measurements. Extended account in Appendix D.

9 Generalization to Non-Clinical Structured Extraction

We repeat the design on a small set of 10 synthetic news articles with schema `{doc_id, events[], persons[], organizations[], locations[]}` — categorically different domain and vocabulary. Tables 10 and 11 show results.

Events collapse completely (F1≈0; mirroring *medications*); *locations* remain stable at F1≈0.71 (mirroring *conditions*). JSON validity and SDS stay at 1.00/0 throughout. The validity–fidelity gap thus appears in both the synthetic clinical setting and a smaller synthetic news setting (10 documents; 1 run), suggesting, but not establishing, that extraction from repeated text in other domains may behave similarly. The news schema uses a simplified binary SDS (Appendix).

10 Implications

Structural monitoring is an incomplete reliability signal. JSON validity and schema-match

Model	5× DRE	10× DRE	16× DRE	Setting
LLaMA-8B	+0.190	+0.204	+0.195	full (1–32×)
LLaMA-70B	+0.008	+0.029	+0.026	midrange repl.
LLaMA-70B (100p)	−0.018	−0.002	+0.005	100 patients
Qwen-72B	+0.012	+0.018	+0.031	midrange repl.

Table 9: **Cross-model DRE** (DRE computed per model). At 8B, repeated evidence is associated with more degradation than filler (large positive DRE). At 70B/72B, DRE ≈ 0: redundancy and filler produce similar F1 trajectories — both *improving* at 70B (Table 8). In all regimes, JSON validity = 1.00 and SDS = 0 (not shown): structural monitoring is uninformative whether F1 is going up or down.

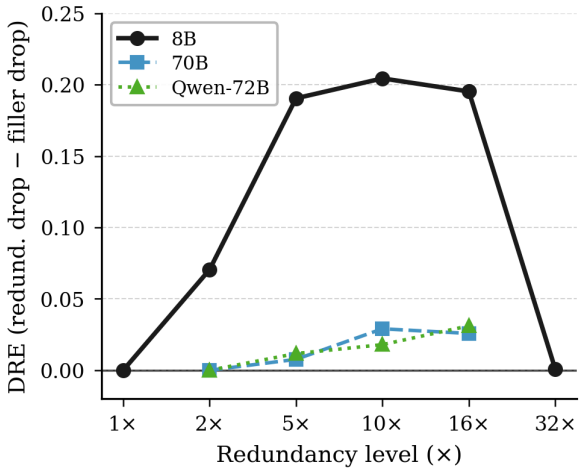


Figure 4: Cross-model DRE. The redundancy-specific component attenuates with scale. Structural validity = 1.00 throughout at all scales (not plotted).

Rep	F1 ↑	FP ↓	FN ↓	JSON valid
1×	0.664	3.1	4.7	1.00
5×	0.605	4.4	5.1	1.00
10×	0.624	4.7	4.7	1.00
16×	0.665	3.9	4.3	1.00

Table 10: News article extraction: redundancy scaling (8B; 10 documents; 1 run). JSON validity remains 1.00 and SDS = 0 throughout even as F1 fluctuates — the validity–fidelity gap also appears in a non-clinical domain. FP and FN are per-document means. F1 dips at 5×–10× before partially recovering at 16×, a weaker but consistent pattern.

Rep	F1 (redund.)	F1 (filler)	DRE
5×	0.605	0.669	+0.062
10×	0.624	0.672	+0.045
16×	0.665	0.660	−0.008

Table 11: News article extraction: DRE (8B). Positive DRE at 5× and 10× is in the same direction as the clinical finding but small and from one run. DRE attenuates to near zero at 16×, where both conditions converge — a weaker effect than the clinical setting (DRE 0.062 vs. 0.190 at 5×), consistent with a smaller and noisier entity inventory in news articles ($n = 10$ documents).

rate are necessary sanity checks but are not sufficient proxies for content fidelity. Our experiments show they remain at their ideal values across all tested conditions — degrading F1, improving F1, and flat F1 alike. A production system that gates processing on structural validity alone could not distinguish high-quality from degraded outputs in these settings.

We recommend augmenting every structured-output pipeline with at least three additional monitoring layers: (1) **semantic monitoring** — entity-level micro-F1, FN and FP counts disaggregated by field type; (2) a **representation-drift indicator** — the fraction of FP strings that contain gold entity content in a non-compliant format (e.g., serialized dict substrings); and (3) a **stress test** at deployment-representative context lengths before any model or pipeline update is shipped. The evaluation protocol in Appendix E operationalizes all three.

Field-type heterogeneity matters. Not all fields are equally at risk. As shown in Figure 1 (§4), high-cardinality fields (medications, observations) reach 50% entity-level persistence loss at 5× redundancy, while conditions — the lowest-cardinality field — never drop below 0.77 of baseline. Practitioners should apply the highest monitoring sensitivity to high-cardinality, richly-structured fields in their schema.

Document-level heterogeneity compounds this: 2 of 10 pilot patients show zero DRE at all levels, and 36–52% of large-N patient cells show zero DRE at individual repetition levels. The aggregate signal is robust (Table 6, CIs exclude zero), but per-document degradation is far from uniform. Relying on aggregate metrics alone will undercount the fraction of documents processed at degraded fidelity.

Constrained decoding: a conceptual argument, not a result. A natural reaction to representation drift is constrained decoding (1rgs, 2023; Willard and Louf, 2023; guidance-ai, 2023). We did not test it. Conceptually, a decoder constrained by the schema used here (array items unspecified) would still permit objects, while one that fixes items to strings would remove the object form we observed; but neither prevents a string such as `"{\"name\":\"acetaminophen\"}"` from being emitted in place of `"acetaminophen 650 mg q6h prn"`. Whether this happens under constraint is an empirical question we leave open; stricter item-level validators (level (3) above) would at least make such cases detectable.

Zero-ground-truth detection. For pipelines without labeled data: flag any call where >5% of array elements contain a JSON brace ({) or are non-string objects (the exact rule is in Appendix A). In the 8B pilot (30 calls per level) this flags 4/30 calls at 1× — all of which are genuine baseline drift (nested-dict medication emissions already present at 1×), not false alarms — rising monotonically to 15/30 at 5× and 24/30 at 32×. *Threshold sensitivity*: detection counts are identical for thresholds between 1% and 20%, because drifted calls emit nested structures for entire fields rather than isolated elements; on these outputs the specific 5% choice is not load-bearing. We deliberately exclude the colon from the character set: legitimate clinical entity strings (e.g., `"Rapid strep test: positive"`) contain colons, a domain-dependent false-positive source, and practitioners should audit their schema's legitimate character distribution before deploying the scan. The heuristic requires no ground-truth labels and deploys as a lightweight post-processing character scan. It does not measure severity — it is a trip-wire, not a metric — and should prompt a labeled evaluation sample when triggered.

11 Conclusion

Representation drift is a change in how entities are written, invisible to structural monitoring in every setting we tested. At 8B, all FN-scored medications we checked are present in the output as nested objects, and a schema-free probe recovers them. At 70B, F1 improves under longer contexts, yet structural validity remains at 1.00. Validity metrics give no signal in either direction.

At 8B the effect is threshold-like (55% of degradation at the 2×→5× step) and is consistent with accumulated semantic content, though section-level, shuffled and paraphrased controls only partly separate this from string form. The effect attenuates with scale. In synthetic settings like ours, a pipeline that monitors only JSON validity could not tell improving from degrading extraction, so we recommend adding entity-level, field-level and drift-indicator checks, with validation on real documents as the next step.

Limitations

- **Synthetic, repeated documents.** All quantitative evidence comes from 10 template-based synthetic notes with exact repetition; real notes and retrieved contexts have heterogeneous, partial overlap and noise, and we make no claim about real-EHR deployment. The "100-patient" sets repeat the 10 templates, so the effective sample size is 10 documents and patient-level CIs are too narrow (Section 6). Real-note replication (requiring IRB/DUA compliance) and retrieval-augmented settings are the main extensions.
- **Small non-clinical study.** The news experiment has 10 documents and one run; it suggests but does not establish generality.
- **Schema dependence.** The prompt schema leaves array items unspecified; "plain string" is our evaluation convention. Results may differ with typed schemas or item-level validators, and only two small schemas were tested.
- **Heuristic length control.** Filler uses four fixed non-clinical sentences sized by a character-count proxy; prompt tokens differ at high repetition (Appendix). DRE is a diagnostic contrast that depends on this baseline, not a causal estimate.
- **Scale, mechanism and scope.** Degradation and large DRE appear at 8B and largely vanish at 70B/72B (32× was not run at 70B), so the robust finding is that validity metrics are uninformative, not that redundancy generally induces drift. Semantic saturation is a hypothesis (no internal measurements). The detection heuristic's threshold and character choice may be schema-dependent. We tested three open-weight models via one provider, and no constrained decoders or mitigations; all evaluators are string-based.

References

1rgs. 2023. Jsonformer: A bulletproof way to generate structured JSON from language models. GitHub repository, https://github.com/1rgs/jsonformer.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Yushi Bai, Xin Lv, Jiajie Zhang, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint*. ArXiv:2308.14508.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint*. ArXiv:2004.05150.

Tom B. Brown and 1 others. 2020. Language models are few-shot learners. *NeurIPS*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. Jsonschemabench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

guidance-ai. 2023. Guidance: A guidance language for controlling large language models. GitHub repository, https://github.com/guidance-ai/guidance.

Cheng-Ping Hsieh and 1 others. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint*. ArXiv:2404.06654.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Greg Kamradt. 2023. Needle in a haystack – pressure testing LLMs. GitHub repository, https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *ACL*.

Nelson F. Liu, Kevin Lin, John Hewitt, and 1 others. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*. ArXiv:2307.03172.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*.

Long Ouyang and 1 others. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Bowen Peng and 1 others. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint*. ArXiv:2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. ArXiv:2302.00093.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*. ArXiv:2104.09864.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US. Association for Computational Linguistics.

Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. 2023. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint arXiv:2309.08963*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint*. ArXiv:2304.10428.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*. ArXiv:2109.01652.

Jason Wei and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*. ArXiv:2309.17453.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. 2025. StructEval: Benchmarking LLMs' capabilities to generate structural outputs. *arXiv preprint arXiv:2505.20139*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Benchmark and Setup (Full Details)

Synthetic documents. 10 synthetic one-page clinical discharge summaries, written as fixed templates in our code in the style of Synthea records (Walonoski et al., 2018) (not produced by the Synthea generator), each with ground-truth JSON under the schema `{patient_id, conditions[], medications[], observations[], procedures[]}`. Sets of "100 patients" are produced by cycling these 10 templates (patient i uses template $((i-1) \bmod 10)+1$), so they contain 10 distinct documents. Notes are rendered to PDF (ReportLab) and the text is extracted with pdfplumber before being sent to the model. The synthetic design enables (i) exact repetition-factor manipulation with all other variables fixed, (ii) unambiguous ground truth, and (iii) entity-string-level FN/FP measurement. The non-clinical set consists of 10 synthetic news articles with schema `{doc_id, events[], persons[], organizations[], locations[]}`.

Redundancy and filler manipulation. For each patient and $r \in \{1, 2, 5, 10, 16, 32\}$, the base note is concatenated verbatim r times. The filler control keeps the note once and appends paragraphs chosen at random (seeded) from four fixed non-clinical sentences (e.g., "The quick brown fox jumps over the lazy dog…") until a character-based estimate $\lceil |x|/4.0 \rceil$ reaches the redundancy condition's length. The filler list is released with the artifacts and was checked post hoc for lexical overlap with gold entity strings; it was not filtered by construction. Backend-reported `usage.prompt_tokens` show substantial mismatches at high repetitions (Table 13); the filler is also far more repetitive and less clinical than the note itself, so DRE depends on this particular baseline and is a diagnostic contrast.

Models and inference settings. All experiments: temperature 0, top-p 1.0, max 512 tokens, seed 1337 (Table 12). Primary pilot: Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024) via DeepInfra. Cross-model: LLaMA-3.1-70B-Instruct (Grattafiori et al., 2024) and Qwen2.5-72B-Instruct (Yang et al., 2024). The dissociation probe used max 256 tokens, timeout 60 s and 3 retries; all other settings are as in Table 12.

Metrics. *Semantic*: micro precision/recall/F1, FN/FP counts (strict, soft and canonicalized variants; definitions in Section 3). *Structural*: raw JSON validity, repair frequency, Structural Drift Score (SDS). For the news schema SDS is simplified to a binary score (1 if the output is malformed or the schema does not match, else 0). *Variance*: run-to-run semantic variability. JSON validity as implemented strips a leading/trailing markdown fence before parsing and requires the result to be a single JSON object.

System prompt.

```
You are a backend medical extraction
service. Return ONLY valid JSON. Do not
use markdown. Do not use code fences.
Do not explain. Do not add commentary.
If information is missing, return empty
arrays. Follow the exact schema.
```

Extraction prompt template (clinical).

```
Extract all medical information from the
document into this EXACT schema:
{"patient_id": "", "conditions": [],
"medications": [],
 "observations": [], "procedures": []}

Rules: Return ONLY valid JSON. Do not
wrap in markdown.
Do not add extra keys. Do not rename
fields.
Arrays must always exist. If empty, use
[].

Document: {document_text}
```

The clinical prompt does *not* state that array elements must be strings; the plain-string representation is implied by the empty-array schema example and by our gold data.

News prompt. The news prompt is analogous (system prompt: "backend news information extraction service"), with schema `{doc_id, events[], persons[], organizations[], locations[]}` and the additional rule `Each array element must be a plain string`.

Dissociation probe prompt.

```
Read the following medical note and list
every medication mentioned. Write one
medication per line. Include the drug
name and dose if present. Do not use
JSON. Do not use any special formatting.
Just plain text.

Document: {document_text}
```

No system prompt is used; the probe is applied to the 1× document for 30 patients (10 templates × 3 copies), and recall is scored against gold medications.

Zero-ground-truth drift check. For each call, count array elements (in the four array fields) that are non-string objects or strings containing a brace character ({). Flag the call if the fraction of such elements exceeds 5%. Colons are deliberately not used because legitimate clinical strings contain them.

Artifact availability. The pipeline, synthetic data, filler list, evaluator and per-call raw model outputs for all reported runs are provided as supplementary material.

Setting	Value
Temperature	0.0
Top-p	1.0
Max tokens	512
Seed	1337
Retries	5 (backoff)
Timeout	120s

Table 12: Fixed inference settings across all experiments.

Inference settings.

Rep	Condition	Prompt tokens (mean±sd)	n
1×	filler	321.0 ± 36.8	100
1×	redundancy	320.1 ± 36.8	100
5×	filler	676.5 ± 107.7	100
5×	redundancy	1032.5 ± 183.8	100
10×	filler	1109.2 ± 198.4	100
10×	redundancy	1923.0 ± 367.5	100
16×	filler	1630.3 ± 308.9	100
16×	redundancy	2991.6 ± 588.1	100

Table 13: Backend-reported `usage.prompt_tokens` for the 70B 100-patient midrange run (repetitions 1/5/10/16; 1 run/condition). This quantifies that the heuristic token-proxy control does not guarantee equal prompt-token counts.

Backend token audit (70B, 100-patient).

Run accounting.

Evaluator. Strict: exact match after case/whitespace/punctuation normalization. Soft: additionally normalizes common clinical variants (HTN/Hypertension, DM2/T2D). Canonical: additionally strips numeric dose/unit/frequency tokens from medication strings. All three variants show the same qualitative trends.

B Error Taxonomy (Full)

Three error types: (1) **Genuine omission**: entity absent entirely (FN, no matching FP).

Experiment	Pts	Calls	Err	Notes
8B redundancy (pilot)	10	180	0	6 reps, 3 runs/cond
8B filler control	10	180	1	matched; 1 repair err
70B replication	10	120	0	6 reps, 1 run
Qwen-72B replication	10	120	0	same
70B midrange (3-run)	10	240	0	reps 2/5/10/16
Qwen-72B midrange	10	240	0	same
8B large-N redundancy	100	400	0	reps 1/5/10/16
8B large-N filler	100	400	0	matched
Shuffled redundancy	10	120	0	reps 5/10/16, 3 runs
Section-level ablation	100	400	0	reps 1/5/10/16
Paraphrased redundancy	50	200	0	reps 1/5/10/16
70B 100-pt redundancy	100	400	0	reps 1/5/10/16
70B 100-pt filler	100	400	0	matched
Prompt robustness	50	300	0	3 prompts, reps 5/10/16
Non-clinical news	10	80	0	reps 1/5/10/16, 2 conds
Dissociation probe	30	30	0	free-form follow-up
Total		**3810**	**1**	

Table 14: Complete run accounting. All runs use temperature 0, fixed seed. Errors are `response_error.txt` markers.

(2) **Representation drift**: entity present in output as a nested object instead of the expected plain string, creating paired FN+FP. E.g., gold `"acetaminophen 650 mg q6h prn"` becomes `{"name":"acetaminophen","dose":"650mg", "route":"po","freq":"q6h"}` in output. (3) **Novel insertion**: FP with no correspondence to any gold entity (hallucination). Drift is distinguished by symmetric FP/FN inflation; novel insertion creates FP only; genuine omission creates FN only. At 1×, medications show a pre-existing format tension (2 TP, 23 FP/FN); drift amplifies this tension — it does not create it.

C Threshold Detail (Per-Patient)

Of 8 patients showing any degradation, 5 show their single largest F1 drop at or before 5×. Three patients (02, 05, 06) show catastrophic drops of 0.35–0.57 F1 at the 2×→5× transition specifically. Two "immune" patients (03, 08) show zero degradation at all levels, indicating document-level factors modulate the threshold. No patient exhibits smooth monotone decline; degradation is always concentrated in one or two transitions.

Strict vs. soft F1.

Canonicalized DRE.

D Scale Account (Extended)

The cross-model pattern supports a **semantic saturation account**: at 8B, format instruction following is overwhelmed by accumulating semantic content load regardless of verba-

Rep	F1 (strict)	F1 (soft)
1×	0.499	0.499
2×	0.457	0.457
5×	0.325	0.333
10×	0.311	0.323
16×	0.292	0.292
32×	0.259	0.259

Table 15: Strict vs. soft micro-F1 (8B). Soft normalization does not change the trend.

Rep	$\Delta F1$ (red.)	$\Delta F1$ (fill.)	DRE
2×	0.043	−0.033	+0.076
5×	0.167	−0.028	+0.195
10×	0.177	−0.028	+0.204
16×	0.204	+0.005	+0.199
32×	0.242	0.242	+0.000

Table 16: Canonicalized DRE (medications normalized to ingredient-like strings). Positive midrange DRE persists; the effect is not explained by dose/frequency surface-form differences.

tim/shuffled/paraphrased form and prompt wording. At 70B/Qwen-72B, the model handles longer contexts more robustly — the 8B pattern is not observed and recall can improve. DRE attenuates because *both* conditions improve similarly, not because both degrade. The phenomenon is scale-dependent in both trigger and direction; structural monitoring is blind in both cases. We treat "semantic saturation" as a descriptive label; mechanism (instruction-token attention dilution under long contexts (Xiao et al., 2024; Shi et al., 2023)) is a hypothesis, not a measured account.

Rep	F1 (exact)	F1 (shuffled)	F1 (filler)
5×	0.325	0.428	0.510
10×	0.311	0.357	0.510
16×	0.292	0.326	0.482

Table 17: Shuffled redundancy control (8B; 10 patients; 3 runs/condition). Breaking exact string identity modestly reduces degradation but does not close the gap to the filler baseline. Both forms of medical repetition perform substantially worse than unrelated filler, confirming the operative stressor is semantic content load, not string form.

Shuffled redundancy (detail).

Paraphrased redundancy (exploratory). A paraphrased-redundancy condition (50 patients; 1 run; reps 1×/5×/10×/16×) yields positive DRE, suggesting representation drift extends beyond exact-string repetition. Caution: a mention-preservation audit finds 48% of gold medication mentions are not recoverable as substrings in generated paraphrases, confounding the comparison. We report this as exploratory and recommend improved paraphrase fidelity verification in future work.

Rep	DRE (verbatim)	DRE (paraphrase)
5×	0.190	0.243
10×	0.204	**0.305**
16×	0.195	0.158

Table 18: Paraphrased vs. verbatim DRE (8B; 50 patients). Paraphrased redundancy does not attenuate drift; DRE at 10× is 0.305 vs. 0.204 verbatim. Caveat: 48% of gold medication strings are not recoverable as substrings in paraphrases, confounding the comparison. The section-level null (Table 7) is the cleaner mechanistic result.

Paraphrased redundancy (detail).

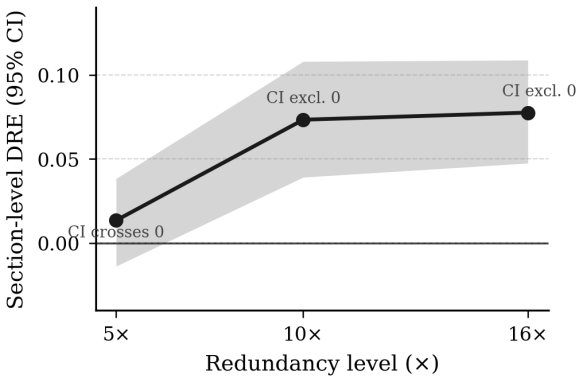


Figure 5: Section-level DRE with 95% bootstrap CI. Section-level redundancy produces smaller but positive DRE at 10×–16×, bridging toward realistic copy-forward EHR patterns.

Section-level ablation.

Prompt	5× DRE	10× DRE	16× DRE
A	0.175	0.144	0.144
B	0.106	−0.033	0.001
C	0.102	0.117	0.059

Table 19: Prompt robustness (50 patients; 1 run; 8B). DRE positive at 5× under all three prompt wordings. Prompt B shows negative DRE at 10×, confirming the measured DRE magnitude is prompt-sensitive while the directional finding is robust.

Prompt-wording robustness (detail).

E Evaluation Protocol (Full)

Recommended evaluation protocol for any structured extraction pipeline:

1. **Structural**: JSON validity and schema-match rate (necessary, not sufficient).
2. **Semantic**: entity-level micro-F1, FN/FP by field type.
3. **Drift indicator**: fraction of FP strings containing gold content in non-compliant format.
4. **Variance**: run-to-run instability at matched conditions.
5. **Stress test**: redundancy/long-context conditions at deployment-representative context lengths.

High-cardinality fields (medications, observations) are earliest indicators (Figure 1; medications reach 50% persistence loss at 5×; conditions never drop below 0.77 of baseline). Document-level heterogeneity: 2/10 pilot patients show zero DRE at all levels; 36–52% of large-N cells show zero DRE depending on level. Notes dominated by medications/observations are at higher risk.

Rep	Mean ΔF1	CI low	CI high
5×	0.156	0.116	0.197
10×	0.170	0.133	0.208
16×	0.190	0.154	0.227

Table 20: Large-N (100 patients, 8B) mean paired F1 drop under redundancy with 95% bootstrap CI (Efron and Tibshirani, 1993). CIs exclude zero; degradation increases monotonically from 5× to 16×.

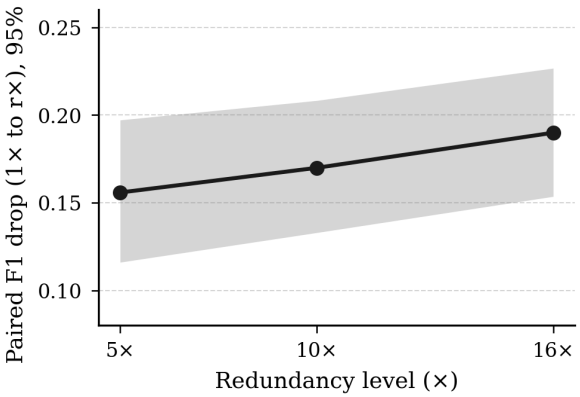


Figure 6: Large-N (100 patients, 8B): mean paired F1 drop under redundancy with 95% bootstrap CI.

Large-N F1 drop under redundancy.

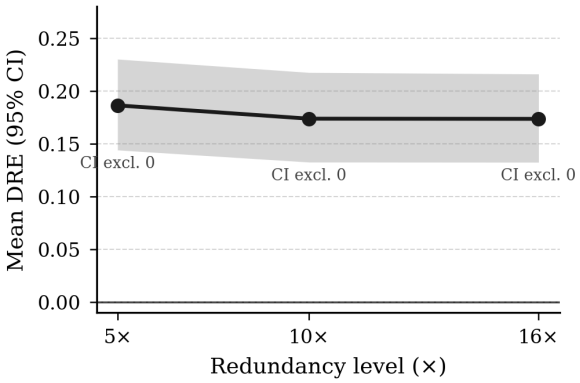


Figure 7: Large-N: mean DRE with 95% bootstrap CI.

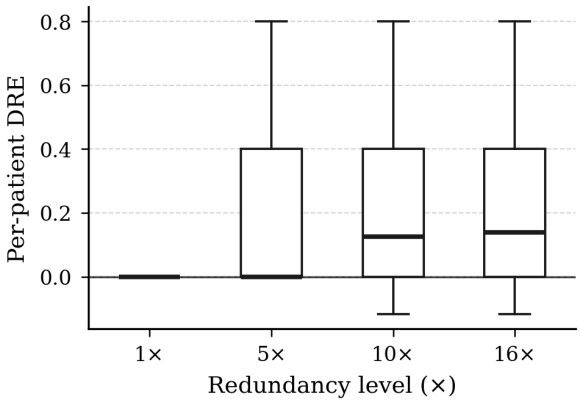


Figure 8: Large-N: per-patient DRE distribution by repetition level.