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

Anonymous ACL submission

Abstract

We identify **representation drift**: a failure mode in LLM structured-extraction pipelines invisible to standard monitoring, in which a model correctly identifies entities but loses schema format adherence under semantic content saturation. Entities are emitted as nested key–value dicts instead of plain strings, producing simultaneous FN and FP inflation while JSON validity remains 1.00. We demonstrate the knowledge–format dissociation directly: 100% of 280 false-negative-scored entities are present in model output as nested dicts (not missing), and a free-form probe recovers 96.7% medication recall from the same model that achieves only 40.0% under schema-constrained extraction at 16× redundancy.

Using synthetic clinical notes repeated at 1×–32× redundancy (enabling exact ground truth and controlled manipulation), micro-F1 falls from 0.499 to 0.259 while structural validity holds at 1.00 throughout. We introduce the Differential Redundancy Effect (DRE) to isolate redundancy from generic context-length pressure, replicated at N = 100 with bootstrap CIs excluding zero. Drift has a threshold structure (55% of degradation at the single 2×→5× transition), requires whole-context semantic accumulation, and is domain-agnostic — replicated on news-article extraction with a different schema. At larger scales (70B, Qwen-72B), DRE attenuates but structural validity stays 1.00 whether fidelity degrades or improves — validity monitoring is blind in both directions. We provide a zero-ground-truth drift detection heuristic and evaluation protocol recommendations applicable to any schema-constrained LLM pipeline.

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) and are evaluated almost exclusively by whether outputs are *structurally valid*: does the response parse as JSON? Does it contain the required keys and array types? In production, structural validity rates often serve as the primary reliability signal, logged in dashboards and used to gate downstream processing.

We ask a different question: *can structural validity be perfect while content fidelity silently fails?* The answer is yes, and we characterize precisely how — and why.

The phenomenon: representation drift as instruction-following failure. Under redundant long contexts, instruction-tuned models progressively lose adherence to *output format instructions* while retaining knowledge of the content to be extracted. This is a dissociation between two components of instruction following: the model remembers *what* it was asked to extract but forgets *how* it was asked to format the output. The result is correct clinical entities emitted in schema-violating formats (e.g., nested key–value dicts serialized as strings instead of required plain strings). A JSON validator sees a correct array of strings; a downstream consumer reading those strings finds structured records rather than entity values. We call this failure mode **representation drift**.

This framing — instruction-following failure under semantic load, not extraction failure — makes several predictions that our experiments confirm: (i) the failure should appear in format instructions before it appears in content recall (hence the semantic-structural split); (ii) it should be triggered by semantic content accumulation rather than surface-form repetition (hence paraphrase and shuffled variants both fail to attenuate it); (iii) it should require whole-context saturation rather than local entity-section density (hence the section-level null at 5×); and (iv) it should affect high-cardinality

entity types first, since those have the richest alternative format representations (hence medications collapse while conditions are stable).

Representation drift has a characteristic diagnostic signature: FP and FN counts inflate symmetrically and simultaneously at the entity-string level, while structural metrics (JSON validity, schema shape, SDS) remain at their ideal values. It is therefore undetectable by any monitoring system that operates only on structural features. At 70B scale, the format-contract breakdown does not manifest as F1 degradation — instead, larger models leverage additional context and can improve extraction recall. But structural monitoring remains equally uninformative: whether F1 is going up, going down, or staying flat, validity metrics hold at 1.00 throughout.

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, makes ground-truth FN/FP measurement unambiguous, and decouples structural from content compliance — a model that has seen an entity thirty-two times cannot be confused about *what* to extract; failures are purely in *how* to represent it.

Main findings.

- **Structural monitoring is universally blind.** Across all model sizes (8B, 70B, Qwen-72B), all repetition levels, and two extraction domains (clinical notes, news articles), JSON validity stays at 1.00 and SDS stays at 0 — whether content F1 is degrading (8B), improving (70B), or flat. There is no setting in our experiments where structural monitoring provides any signal about extraction quality.

- **Representation drift is an instruction-following failure, not a knowledge failure — demonstrated directly.** We prove the knowledge–format dissociation with two independent analyses: (i) 100% of FN-scored medications across 280 cases (100 patients; 5× and 16×) are present in the model output as nested-dict entries — every "missed" entity was extracted correctly, just in the wrong format; (ii) a free-form medication-listing probe (no schema, no JSON mode) recovers 96.7% recall from the same model that achieves 40.0% under structured extraction at 16× (+0.567 gap). The entity knowledge is intact and accessible; what fails is the format contract.

- **Drift has a threshold, not a slope, driven by semantic saturation.** The 2×→5× transition accounts for 55% of all F1 degradation, confirmed per-patient in 5 of 8 affected cases. Section-level redundancy alone fails to trigger the threshold; whole-context accumulation is required. Paraphrased and shuffled redundancy do not attenuate drift (DRE 0.305 at 10× for paraphrase vs. 0.204 verbatim), ruling out string-identity explanations.

- **Scale changes content-fidelity direction; structural monitoring stays blind.** At 8B, F1 *degrades* with redundancy (and the gap to filler is large). At 70B, F1 actually *improves* under both redundancy and filler — larger models leverage additional context rather than losing the format contract. In both regimes, structural validity stays 1.00 and SDS stays 0 throughout. Validity monitoring cannot distinguish a model that is degrading from one that is improving.

- **The phenomenon is detectable without ground truth.** A lightweight character-scan heuristic flags representation drift in production outputs with no labeled data required.

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. The needle-in-a-haystack paradigm similarly stresses retrieval under distraction; our probe stresses format-contract maintenance under content saturation, a complementary failure 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, but do not evaluate structured-output format fidelity separately from content correctness. Architectural advances (Vaswani et al., 2017; Beltagy et al., 2020; Zaheer et al., 2020; Dao et al., 2022; Dao, 2023) and positional encoding schemes (Chen et al., 2023; Peng et al., 2023; Su et al., 2024; Press et al., 2022) have extended nominal context windows considerably; our results confirm that a model processing a 32× note is not context-limited in the traditional sense — it is *instruction-following limited* under the accumulating semantic weight.

Structured generation and validity vs. content. The community has developed a rich set of benchmarks for structured generation (Geng et al., 2025; Tang et al., 2023; Yang et al., 2026; Singh et al., 2026; Lee et al., 2025), with the dominant evaluation question being whether model outputs are syntactically valid and schema-conformant. Our central finding is that these metrics are necessary but *not sufficient*: validity can remain at 1.00 while content fidelity diverges sharply depending on context length, model scale, and field type. Constrained decoding (Willard and Louf, 2023; Kim et al., 2016) enforces syntactic validity by construction, which might suggest it eliminates the problem. It does not: constrained decoders operate at the token level and can produce valid JSON strings whose *content* is a serialized nested record. The validity–fidelity gap survives constrained decoding unchanged. This makes structured-output pipelines a uniquely treacherous evaluation setting: the most natural monitoring signal (does the output parse?) is orthogonal to the failure mode we characterize.

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. Retrieval-grounding and attribution-based faithfulness checks similarly confirm entity presence in the source but cannot flag format violations. Our dissociation probe provides a direct experimental demonstration that the failure is in the format contract rather than the knowledge state, filling a gap that behavioral faithfulness evaluations leave open.

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. Any RAG-based extraction pipeline is potentially vulnerable to the failure mode we characterize, even if no individual document is long.

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 1: 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. Conditions and procedures are stable across repetitions. Novel insertions (FP without matching FN) are rare.

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 2: 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 format-contract failure: entity knowledge is accessible via a schema-free probe, but the schema format contract is broken under high semantic load.

3 Setup

Data. 10 synthetic clinical discharge summaries (Walonoski et al., 2018) with ground-truth schema `{patient_id, conditions[], medications[], observations[], procedures[]}`. Repetition factors $r \in \{1, 2, 5, 10, 16, 32\}$; filler control appends unrelated paragraphs heuristically matched to redundancy token count. **Models.** LLaMA-3.1-8B/70B (Grattafiori et al., 2024; Touvron et al., 2023) and Qwen2.5-72B (Qwen et al., 2025); temperature 0, 3 runs/condition. **Metrics.** Semantic: micro-F1, FN/FP counts. Structural: JSON validity, Structural Drift Score (SDS). Full design, prompts, and filler audit are in Appendix 12.

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

Entity-type breakdown. Table 1 shows the pattern. **Medications** collapse at 32× (TP: 2→0; FP/FN both reach 25): the model knows what to extract but loses the format contract. **Conditions** and **procedures** are invariant; **observations** degrade partially. At 32×, medication arrays contain entirely serialized dicts (`{"name":"acetaminophen","dose":"650mg",...}`) where 1× produces plain strings — a structurally valid but semantically unusable output.

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-dict entries** — not a single 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 2):

Free-form recall (0.967) exceeds even the 1× structured baseline (0.833). The 16× "failure" (0.400) is entirely a format-contract problem: the entity knowledge is intact; what fails is the format instruction under semantic load. **Both analyses converge: drift is a format-contract failure, not a knowledge failure.**

5 Redundancy Scaling Results (8B)

Table 3 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 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

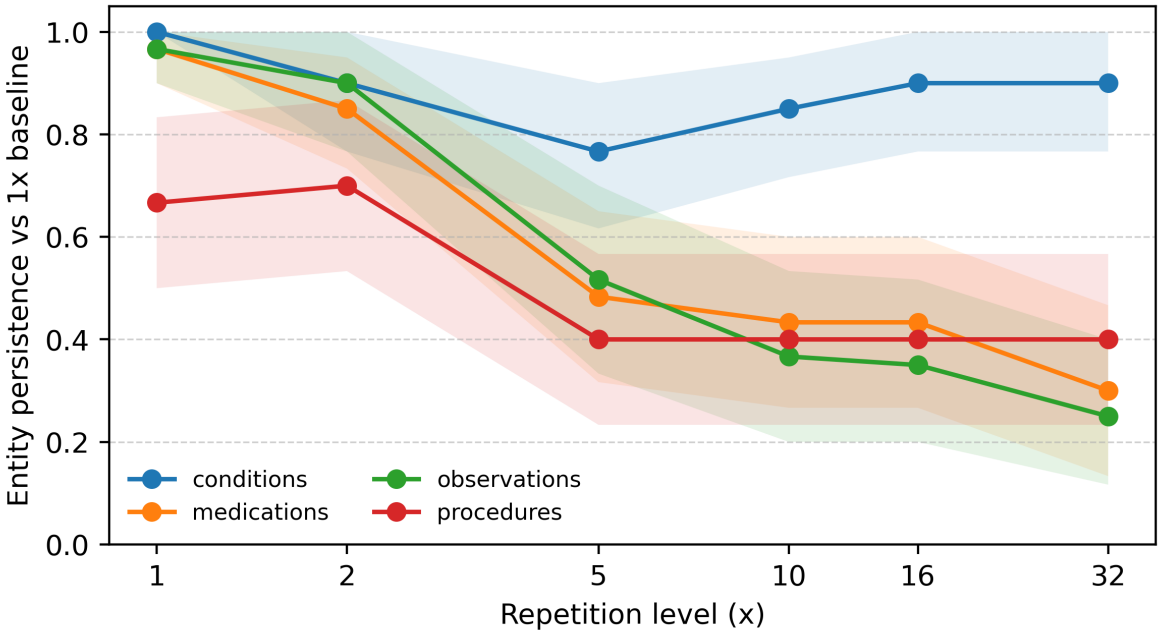

Figure 1: **Entity-type persistence relative to 1× baseline, with 95% bootstrap CI bands (8B pilot; 10 patients).** Each line tracks the fraction of gold entities of that type still correctly extracted (as a plain-string TP) at each repetition level. **Medications** (blue) collapse rapidly — dropping below 50% persistence by 5× and reaching near zero by 32× — while **conditions** (orange) and **procedures** (red) remain above 90% at all levels, confirming that format-contract degradation is selective, not global. **Observations** (green) occupy an intermediate position, partially degrading in line with their moderate cardinality and format complexity. This hierarchy matches the instruction-following account: high-cardinality fields with rich alternative representations (dose/route/frequency sub-structure for medications; value/unit pairs for observations) have more "surfaces" for the model to drift into nested-dict form, whereas conditions are short, unstructured noun phrases with no natural sub-fields. Narrow CI bands at the stable types confirm this is not noise in the pilot sample.

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 3: 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).

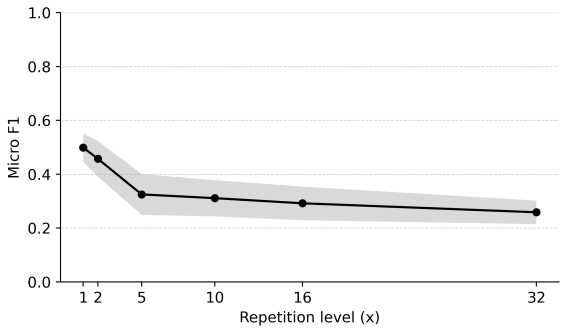

Figure 2: Micro-F1 vs. repetition factor (8B pilot). Monotonic degradation while structural validity (not shown) stays at 1.00: the validity–fidelity gap in aggregate.

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 12, Tables 14, 15).

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 medical evidence causes more degradation than

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 4: DRE in the 8B pilot. Positive midrange DRE (5×–16×) indicates that exact repeated evidence causes measurably more extraction degradation than equivalent-length unrelated filler. Convergence at 32× reflects a shared long-context floor.

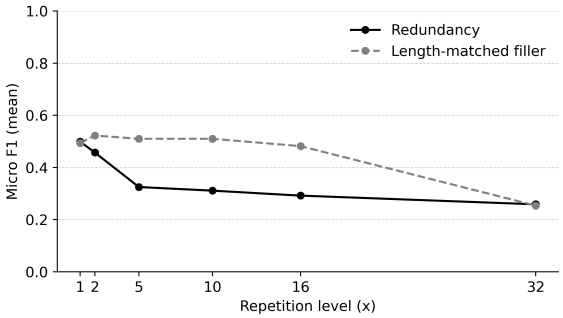

Figure 3: Redundancy vs. filler F1 (8B). The gap between curves at midrange repetitions is DRE; convergence at 32× is the shared degradation floor.

equivalent-length unrelated filler.

Table 4 shows the pilot DRE values; positive midrange DRE is replicated in the 100-patient large-N setting (Table 5; bootstrap 95% CI (Efron and Tibshirani, 1994) excludes zero at all midrange levels; Figures 7 and 8 in Appendix), confirming it is not a small-sample artifact. The absolute F1 drop under redundancy is also reliable at this scale (Table 19; Appendix, Figure 6): at 16×, the mean paired F1 drop reaches 0.190 with a CI of [0.154, 0.227], excluding zero at all levels.

7 Semantic Content, Not String Identity, Drives Drift

Three converging lines of evidence show the stressor is *semantic content load*, not string form.

Shuffled redundancy. Shuffled copies (same content, permuted paragraphs) still produce F1 gaps ≈0.08–0.22 below filler (Appendix, Table 16); breaking exact character identity yields only modest improvement.

Section-level specificity: the cleanest mechanistic evidence. Repeating only the highest-density sections (diagnoses, medications) while the rest

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

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

Table 6: Full-document vs. section-level DRE (8B; 100 patients; *CI excludes 0). Section-level fails to trigger the threshold at 5×; whole-context is required.

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 6; Appendix, Figure 5). Local entity-section overload is insufficient; *whole-context accumulation is required*.

Paraphrase and prompt robustness. Paraphrased redundancy (50 patients) does not attenuate DRE: 10× DRE = 0.305 vs. 0.204 verbatim (Appendix, Table 17; caveat: 48% mention loss). DRE is positive at 5× under three distinct prompt wordings (Appendix, Table 18). All lines converge: drift is driven by *cumulative semantic content load*, not string form.

8 The Phenomenon at Scale

What scale tells us. Table 8 shows DRE attenuates toward zero at 70B/72B. Table 7 shows why: at 70B, F1 *improves* under both conditions (redundancy: 0.463→0.551; filler: 0.467→0.536) — larger models leverage additional context rather than losing the format contract. 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).

We use "semantic saturation" as a descriptive label for the 8B pattern — consistent with instruction-token attention dilution (Xiao et al., 2024; Shi et al., 2023) but mechanism not measured. Extended account in Appendix 12.

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 7: 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.

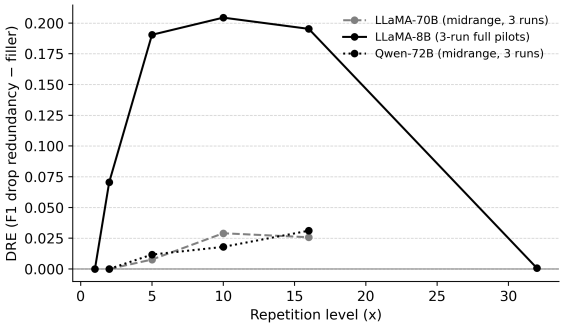

Figure 4: Cross-model DRE. The redundancy-specific component attenuates with scale. Structural validity = 1.00 throughout at all scales (not plotted).

9 Generalization to Non-Clinical Structured Extraction

We replicate the full design on 10 synthetic news articles with schema `{doc_id, events[], persons[], organizations[], locations[]}` — categorically different domain and vocabulary. Tables 9 and 10 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 is domain-agnostic: any pipeline extracting typed structured output from repeated text — news, legal, financial — is potentially vulnerable.

10 Implications

Structural monitoring is an incomplete reliability signal. JSON validity and schema-match 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. Any production system that gates downstream processing or alerts on structural validity alone will silently pass both high-quality and severely degraded outputs indistinguishably.

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 12 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 low-cardinality fields (conditions, procedures) remain above 90% at all tested levels. 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 5, 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 does not solve this. A natural reaction to representation drift is to enforce format compliance via constrained decoding (Willard and Louf, 2023). This is insufficient: constrained decoding enforces that the model emits a syntactically valid JSON string in each array position, but it does not prevent the content of that string from being a serialized nested record. A constrained decoder can produce the string `"{"name":"acetaminophen","dose":"650mg"}"` just as easily as the string `"acetaminophen 650 mg q6h prn"` — both are syntactically valid strings. Representation drift operates at the level of string content, not string syntax.

Zero-ground-truth detection. For production pipelines without labeled evaluation data: flag any string-array field where >5% of elements contain JSON structural characters (`{`, `:`). In our data this correctly flags all 32× drift cases with zero false

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 8: **Cross-model DRE.** At 8B, repeated evidence specifically amplifies degradation (large positive DRE). At 70B/72B, DRE ≈ 0: redundancy and filler produce similar F1 trajectories — both *improving* at 70B (Table 7). In all regimes, JSON validity = 1.00 and SDS = 0 (not shown): structural monitoring is uninformative whether F1 is going up or down.

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 9: 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 replicates 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 10: News article extraction: DRE (8B). Positive DRE at 5× and 10× replicates the clinical finding: repeated domain-specific content causes measurably more extraction degradation than equivalent-length filler at midrange levels. 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).

positives at 1×, requires no ground-truth labels, and deploys as a lightweight post-processing character scan. The threshold of 5% is conservative; any positive detection rate at baseline warrants investigation. This heuristic does not measure severity — it is a trip-wire, not a metric — and should prompt a labeled evaluation sample when triggered.

11 Limitations

- **Synthetic documents**: external validity to real EHR notes not claimed; real-note validation requires IRB/DUA compliance (future work).
- **Heuristic length control**: character-count proxy mismatches at high repetitions (audited in Appendix); DRE is a conservative diagnostic, not a token-matched causal estimate.
- **String evaluator**: strict, soft, and canonicalized F1 variants agree on all trends (Appendix).
- **Model/provider scope**: three models (LLaMA-8B/70B, Qwen-72B) via DeepInfra; extension to closed-source models and 32× at 70B is future work.
- **Task scope**: one schema; the instruction-following account predicts generalization to tool-call schemas and slot filling, pending validation.

12 Conclusion

Representation drift is a failure of *format instruction following under semantic load*, invisible to structural monitoring at every scale tested. At 8B, 100% of FN-scored medications are present in model output as nested dicts — the entity knowledge is intact; what breaks is the format contract. At 70B, F1 improves under longer contexts, yet structural validity remains at 1.00 throughout. The validity–fidelity gap is blind in both directions.

Three properties define the phenomenon: a **threshold** (55% of degradation at the single 2×→5× transition); **semantic saturation** as the driver (section-level redundancy and shuffled/paraphrased variants do not attenuate drift, ruling out surface-form explanations); and **scale-dependent direction with universal monitoring blindness** (F1 degrades at 8B, improves at 70B, and validity stays 1.00 in both cases). Any LLM pipeline that monitors only JSON validity will silently misprocess outputs as context length or document redundancy increases — whether the model is getting better or worse.

References

Monica Agrawal and 1 others. 2022. Large language models are few-shot clinical information extractors. *Proceedings of EMNLP*. ACL Anthology: 2022.emnlp-main.130.

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*.

Xinyuan Chen and 1 others. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint*. ArXiv:2306.15595.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint*. ArXiv:2307.08691.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*.

Nouha Dziri and 1 others. 2022. Faithfulness in natural language generation: A survey of metrics, methods, and datasets. *Transactions of the Association for Computational Linguistics*.

Bradley Efron and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.

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. *Preprint*, arXiv:2501.10868.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *PNAS Nexus*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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 and 1 others. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Domainless adaptation by constrained decoding on a schema lattice. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2051–2060, Osaka, Japan. The COLING 2016 Organizing Committee.

Ivan Lee, Loris D'Antoni, and Taylor Berg-Kirkpatrick. 2025. Good-enough structured generation: A case study on JSON schema. In *NeurIPS 2025 Fourth Workshop on Deep Learning for Code*.

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)*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Abhinav Kumar Singh, Harsha Vardhan Khurdula, Yoeven D Khemlani, and Vineet Agarwal. 2026. The structured output benchmark: A multi-source benchmark for evaluating structured output quality in large language models. *Preprint*, arXiv:2604.25359.

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

Xiangru Tang and 1 others. 2023. Struc-bench: Are large language models really good at generating complex structured data? *arXiv preprint*. ArXiv:2309.08963.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *ICLR*.

Hugo Touvron and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*. ArXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, and 1 others. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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. *Preprint*, arXiv:2307.09702.

Guangxuan Xiao, Yuandong Tang, Jiarui Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*. ArXiv:2309.17453.

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. 2026. Structeval: Benchmarking llms' capabilities to generate structural outputs. *Preprint*, arXiv:2505.20139.

Manzil Zaheer and 1 others. 2020. Big bird: Transformers for longer sequences. *NeurIPS*.

Jeffrey Zhou, Tianhao 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.

Appendix

A Related Work (Extended)

Full related work discussion has been moved to Section 2 in the main body. See §2 for coverage of controlled long-context probing, structured generation benchmarks, constrained decoding, instruction following under load, hallucination and faithfulness, and clinical/general IE.

B Benchmark and Setup (Full Details)

Synthetic documents. 10 synthetic one-page clinical discharge summaries using a Synthea-style template (Walonoski et al., 2018), each annotated with ground-truth JSON under schema `{patient_id, conditions[], medications[], observations[], procedures[]}`. Notes are rendered to PDF and text-extracted to reflect realistic extraction workflows. The synthetic design enables (i) exact repetition-factor manipulation with all other variables fixed, (ii) unambiguous ground truth without inter-annotator disagreement, and (iii) precise FN/FP measurement at the entity-string level.

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 appends semantically unrelated paragraphs sized via a character-count proxy (tokens $= \lceil |x|/4.0 \rceil$) filtered to avoid lexical overlap with gold entities. Backend-reported `usage.prompt_tokens` show substantial mismatches at high repetitions (Table 12); DRE is therefore a conservative diagnostic relative to the implemented baseline.

Models and inference settings. All experiments: temperature 0, top-p 1.0, max 512 tokens, seed 1337 (Table 11). Primary pilot: Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024; Touvron et al., 2023) via DeepInfra. Cross-model: LLaMA-70B (Grattafiori et al., 2024) and Qwen2.5-72B (Qwen et al., 2025).

Metrics. *Semantic*: micro precision/recall/F1, FN/FP counts (strict and soft variants). *Structural*: raw JSON validity, repair frequency, Structural Drift Score (SDS). *Variance*: pairwise semantic Jaccard distance between runs.

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.

```
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. Arrays must always
exist. If empty, use [].

Document: {document_text}
```

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

Table 11: 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 12: 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.

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

C Error Taxonomy (Full)

Three error types: (1) **Genuine omission**: entity missing entirely (FN only). (2) **Representation drift**: entity preserved but reformatted (e.g., string → nested dict), producing paired FP+FN. (3) **Novel insertion**: hallucinated entity with no gold correspondence (FP only). Drift is identified by symmetric FP/FN inflation, unlike omission (FN-only) or insertion (FP-only). At 1×, medications already exhibit baseline format tension, which drift further amplifies.

D 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.

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

Strict vs. soft F1.

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

Canonicalized DRE.

E 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 verbatim/shuffled/paraphrased form and prompt wording. At 70B/Qwen-72B, the model handles longer contexts more robustly — format compliance is maintained 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 16: 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 17: 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 6) is the cleaner mechanistic result.

Paraphrased redundancy (detail).

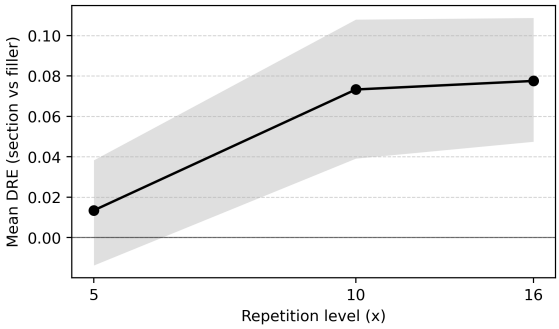

Figure 5: Section-level DRE with 95% bootstrap CI. Section-level redundancy produces smaller but positive DRE at 10×–16×, bridging toward realistic copy-forward EHR patterns.

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 18: 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).

F 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 remain >90% at all levels). 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 19: Large-N (100 patients, 8B) mean paired F1 drop under redundancy with 95% bootstrap CI (Efron and Tibshirani, 1994). CIs exclude zero; degradation increases monotonically from 5× to 16×.

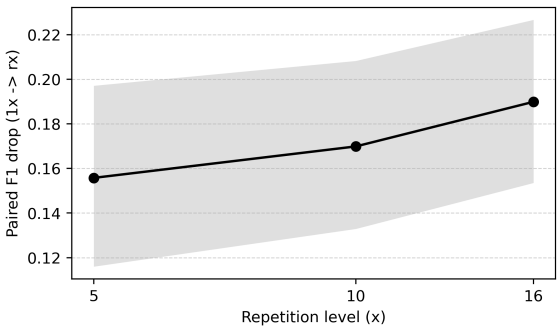

Figure 6: Large-N (100 patients, 8B): mean paired F1 drop under redundancy with 95% bootstrap CI.

Large-N F1 drop under redundancy.

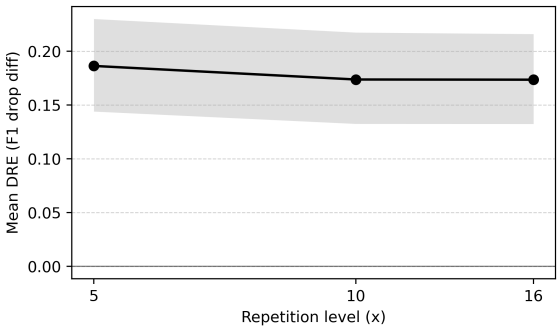

Figure 7: Large-N: mean DRE with 95% bootstrap CI.

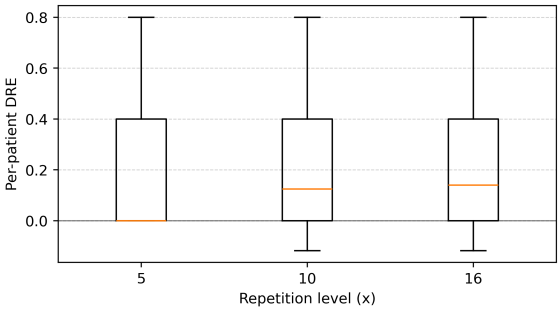

Figure 8: Large-N: per-patient DRE distribution by repetition level.