Failure Modes of Imperfect Relational Context in LLM Document Extraction

Boris Kundu
Independent Researcher
boriskundu@gmail.com

Pooja Mehta
Independent Researcher
poojamehta.mba@gmail.com

Abstract

LLM-based extraction pipelines increasingly consume relational context produced upstream, for example by a heuristic linker, a retrieval step, or another model. We study how imperfect relations between document entities affect LLM extraction on FUNSD, a benchmark of scanned forms. Using identical instructions in every condition, four LLMs receive a form's entities with no relations (Flat), relations from a heuristic linker (Heuristic), gold relations (Oracle), or relations that are almost all wrong (Deranged). In a secondary contrast, gold relations raise accuracy for every model (4.0 to 7.9 percentage points, pp). Heuristic and deranged relations lower accuracy by 8 to 25 pp for three models, while GPT-5 shows significantly smaller effects. Comparing the same questions across conditions shows what lies behind the drops. Wrong edges are associated with answers copied from a different field, and missing edges with more blank answers even though the answer text is still present. These patterns appear under two prompt variants, but the prompts were developed on this corpus and every result is a single generation, so we treat the prompt comparison as exploratory.

1 Introduction

LLM-based systems increasingly consume structured context they did not generate themselves, such as retrieved passages, a knowledge graph neighborhood, or, as we study here, a relation graph over a document's own entities produced by an upstream linker. This context is supplied to change what the system produces, but it is rarely perfect, and a pipeline has no principled way to know how much to rely on a relation it did not itself verify. The same issue arises whenever an agent is grounded in a knowledge graph that an upstream component built, since its edges can be wrong or missing.

We study one concrete instance of this interface. Document pipelines often run OCR and layout analysis first and then hand an LLM typed text spans, their positions, and links between them. We supply this kind of input, with gold segmentation, in a form the model reads as text, and vary only the links. We choose it because it lets us control and score which relations the model sees. We do not test alternatives such as the page image or layout-derived HTML or Markdown (Limitations item 3), so our findings concern this interface, not the best way to prompt for extraction. That correct relations help and wrong ones hurt is expected. What the controlled setting adds is how the failures differ by type of relation error and by model.

Most prior work asks a different question. Models with a graph layer over document entities are trained end-to-end and scored on a fixed benchmark (Lee et al., 2022; Gemelli et al., 2022; Zhang et al., 2023; Hui et al., 2025), which tests whether an architecture helps, not whether the output of an existing general-purpose LLM changes when it is handed the same relations as input context. Work on LLMs consuming graph-structured prompts studies input format, serialization, and, in one case, robustness to noisy or incomplete subgraphs (Dai et al., 2025; Sun et al., 2025), but not inside a document-extraction task that fixes entity text, labels, and layout and decomposes behavior by whether a relation is correct, wrong, or absent.

We hold the extraction task, pipeline, and document fixed and vary only the relations in the input. We build four representations of every FUNSD document (Jaume et al., 2019) that carries a scorable relation: flat entities with type labels and layout but no relations, a relation graph built by a layout-proximity heuristic over gold[1] entities, the gold relation graph, and a deranged graph in which

[1]Here and throughout, "gold" denotes ground-truth values supplied as *input* to the model (entity text, type labels, boxes, or relations); it is distinct from the gold *answer* key used only for scoring (Section 3.5), which the model never sees.

Proceedings of the 3rd Workshop on KNOwledge GRaphs & Agentic Systems Interplay (NORA 2026), pages 26–38
November 10, 2026 ©2026 Association for Computational Linguistics

nearly every edge is wrong by construction. A fixed extract-then-verify LLM pipeline answers every question field and then re-reads the document to revise it, for four models from four labs (Claude Sonnet 5, GPT-5, Qwen3.6-27B, Gemini 3.7 Flash), under one prompt that is identical across conditions and checked programmatically. Two preliminary rounds on this corpus used other wordings and each led to a different, partly wording-driven conclusion (Appendix A); we treat only the third, fully condition-invariant round as primary evidence.

Our contributions are: (1) a controlled document-extraction comparison in which entity text, labels, and layout are held fixed while the supplied relational channel varies; (2) a matched question-level decomposition identifying distinct error signatures for correct, wrong, and absent heuristic relations; (3) evidence that the effects of heuristic and deranged relational context differ substantially across the tested models, with GPT-5 significantly less affected than the other three, a model-dependence result and not a universal LLM behavior. An exploratory cross-prompt sensitivity analysis on the preliminary rounds supplements these findings (Section 4.3). Figure 1 and Table 2 illustrate the setup on a small example, and code, prompts, and aggregate analysis files are available at `https://github.com/boriskundu/relational-context-failure-modes`.

2 Related Work

Graph-structured models for form understanding. FormNet (Lee et al., 2022), Doc2Graph (Gemelli et al., 2022), GraphDoc (Zhang et al., 2023), and SEG-Doc (Hui et al., 2025) each add an explicit graph layer, convolution, attention, or a text-histogram graph, over document entities to recover layout structure lost during serialization. RE^2 (Ramu et al., 2024) targets the same FUNSD entity-linking task we use to build our heuristic graph condition. All of these are trained end-to-end on FUNSD or a similar benchmark, so their graph layer is optimized on the structure it consumes. To avoid a learned relation component fitted to our own evaluation documents, we use an untrained layout-proximity heuristic and report its precision, recall, and F1 directly.

LLMs consuming graph-structured prompts. Dai et al. (2025) study how LLMs process knowledge-graph facts under different linearizations, from raw triples to fluent text, finding systematic format preferences at both the output and attention level, and additionally report that larger LLMs are more susceptible to noisy, incomplete subgraphs. Sun et al. (2025) introduce a structured-prompt benchmark showing general-purpose LLMs with the right graph-prompt design can match specialized graph models on general graph-learning tasks. Our setting instead varies only the supplied key-value relations, with entity text, labels, and layout fixed, and decomposes behavior by whether a given heuristic relation is correct, wrong, or absent.

LLM-based, prompt-driven document extraction. Closer to our own setup, Colakoglu et al. (2025) map the design space of applying general-purpose LLMs to layout-rich document extraction directly, varying input representation, prompting, and model choice, and show a well-configured LLM can match specialized extractors. Their design space includes representation as a factor but does not isolate relational structure or its correctness as an independent variable, our narrower focus. Other multi-stage document pipelines, ontology-guided extraction with human-in-the-loop validation (Chhetri et al., 2025) and a five-agent enterprise pipeline evaluated on automation rate (Gosmar and Zenezini, 2026), likewise do not vary relational structure's quality as a controlled variable.

LLM robustness to conflicting or noisy context. Work on how LLMs handle context that conflicts with what they already know or with itself (Xu et al., 2024) shows that models can be disproportionately influenced by misleading context that resembles relevant evidence, including that retrieval-augmented models are disproportionately misled by counterfactual noise resembling correct context (Hong et al., 2024) and that LLMs can favor plausible-looking self-generated context over correct retrieved context (Tan et al., 2024). That work studies free-text conflicting context whose correctness is not independently known; we instead supply a structured relation channel whose ground-truth relation status is known, letting us decompose a model's response by whether the specific relation it received was correct, wrong, or absent, rather than by how the conflicting context was generated. Graph-retrieval-augmented generation (Zhang et al., 2025) studies a related but distinct question, how a system should retrieve and traverse a graph at query time, whereas our structure is supplied directly in context.

Condition	Text	Labels	Boxes	Relations
Flat	✓	gold	gold	
Heuristic graph	✓	gold	gold	heuristic
Deranged graph	✓	gold	gold	wrong
Oracle graph	✓	gold	gold	gold

Table 1: Primary input representations. Entity text and gold labels are fixed across conditions.

3 Method

3.1 Dataset

We use FUNSD (Jaume et al., 2019), 199 real, noisy scanned forms with human-verified entity text, type labels, bounding boxes, and an entity-linking annotation connecting questions to answers. We use the original annotation release, not the later revision (Vu and Nguyen, 2020), whose files required gated cloud-storage access we did not have; FUNSD's known annotation-quality issues are a limitation (see Limitations). Five of the 199 documents carry no scorable question-answer link and contribute nothing under any condition or model; every reported accuracy is a field-level average over the remaining 194, with model- and condition-specific exclusions in Limitations item 1. A meaningful fraction of questions link to two or more answer entities, usually one logical answer wrapped across two lines; we concatenate all linked entities per question, in reading order, into one scored answer, a disclosed simplification.

3.2 Four representation conditions

Every document is rendered with entity text held constant across all conditions and layout (bounding boxes) held constant across flat and every graph condition. Table 1 summarizes the four conditions, and Table 2 shows them on a small constructed form. In every condition the model sees a list of spans, each with its text, a label (question, answer, header, or other), and its box. The conditions differ only in whether a list of links between spans is added, and which links. **Flat** adds none. **Heuristic** adds the links produced by a layout-proximity rule. **Oracle** adds FUNSD's gold links. **Deranged** adds a list of the same size and degree structure as the gold one but with the answers reassigned so that nearly every link points to the wrong answer. A background raw condition (text only) is scored separately against flat (Appendix C).

The heuristic graph's edges come from a layout-proximity rule, matching each answer to the nearest question above or to its left, with tie-breaking on remaining distance. We chose this over an LLM-inferred relation (which would entangle any benefit with the model's own self-consistency) or an existing FUNSD-trained linker (which risks the circularity noted above). We measure the heuristic's own precision, recall, and F1 on two disjoint parts of the full 199-document corpus, a 35-document validation subset (precision 0.53, recall 0.43, F1 0.47, Appendix E gives the raw edge counts) and its 164-document complement (precision 0.50, recall 0.45, F1 0.47), a substantially imperfect relation source.

The deranged graph replaces every edge with a wrong one through a degree-preserving derangement over entity identity, not a naive random shuffle, which can leave some edges correct by chance. A few resulting edges are still identity-correct or point at text-duplicate answers; these are disclosed and excluded from any diagnostic that requires a genuinely wrong, text-distinguishable edge (Appendix E gives exact counts). Five documents, each with only one or two scorable questions of a within-document-unique multi-value degree, have no same-degree partner to swap with and cannot be deranged, so every model's Deranged condition excludes them by design (Limitations item 1 gives document counts).

3.3 Pipeline and models

Each document, condition, and model cell is processed by the same fixed extract-then-verify LLM pipeline (a 2-node LangGraph implementation, not an autonomously planning agent). An *extract* step produces an initial answer for every question field, and a forced *verify* step re-reads the same representation and corrects any answer it finds ungrounded or wrong; both the initial and final answers are retained and scored. We use four models from four labs: Claude Sonnet 5 (Anthropic), GPT-5 (OpenAI), Qwen3.6-27B (open-weight, served via Groq), and Gemini 3.7 Flash (Google), each fixed to a low or default reasoning setting that keeps reasoning enabled (Table 5; Groq exposes no lower enabled setting for Qwen3.6-27B), held constant across all conditions. We chose them for diversity of provider and openness, API availability, and a reasoning mode we could bound in cost and latency, not for matched size or architecture (undisclosed by the three proprietary providers), so differences between models are descriptive and we make no claim about why they arise. A fifth candidate was dropped before any evaluation because its reason-

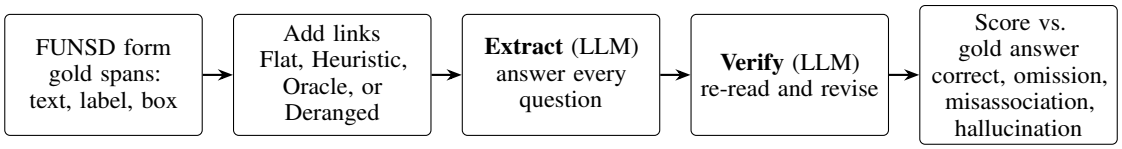


Figure 1: Experimental pipeline. Only the second box varies across conditions. The instruction text, model, and scoring are identical, and both the initial (Extract) and final (Verify) answers are scored.

Condition	Links shown	Resulting bucket for each question
Answer key	(not shown) Date → March 4; Name → J. Smith; Brand → Acme	
Flat	none	no links shown
Heuristic	Date → March 4; Name → Acme	Date: correct edge; Name: wrong edge; Brand: no edge
Oracle	Date → March 4; Name → J. Smith; Brand → Acme	all correct edges
Deranged	Date → J. Smith; Name → Acme; Brand → March 4	all wrong edges

Table 2: A constructed form (not taken from FUNSD) with six spans, three questions (Date, Name, Brand) and their answers (March 4, J. Smith, Acme). In the Heuristic row the three questions fall into the correct-, wrong-, and absent-edge buckets of Section 4.2. Answering "Acme" for Name there is a misassociation (a real value from the Brand field), and answering blank for Brand, although "Acme" is in the input, is a blank response.

ing length could not be bounded.

3.4 Prompt design

Every condition receives one shared instruction block that differs only in the inserted document representation. Prompt identity outside that substitution is checked programmatically, alongside a banned-word list for the input-specific paragraph and a check that the instruction text is byte-identical across the four conditions above plus Raw (Appendix A gives the full safeguard list). The wording was finalized after two preliminary rounds on the same corpus. Round 1 confounded relational structure with an instruction to treat shown links skeptically, and Round 2 removed that instruction but inadvertently also a needed link-explanation sentence, after which two models showed a severe blank-refusal pattern on questions with no shown heuristic edge. Round 3 restores one short, neutral sentence explaining that a link, when present, associates a question with an answer, and that some questions have none; restoring it did not reliably reduce that pattern (Section 4.3). All primary results in Section 4 use Round 3 alone. The two failure signatures of Section 4.2 persist under both condition-invariant regimes, Round 2 and Round 3, which we treat as exploratory sensitivity evidence rather than validation, since both were developed on this same corpus. Appendix A gives the full prompt history and literal templates.

3.5 Scoring

Each predicted answer is scored against the gold answer key with an exact normalized match (lowercased, whitespace collapsed) and a fuzzy similarity score (rapidfuzz `fuzz.ratio`, a normalized edit-based similarity from 0 to 100, with a match at 85 or above). The fuzzy match is tolerant of formatting and OCR-level differences and is our accuracy measure throughout. We fixed the threshold of 85 at the start of development, before any evaluation run, and never tuned it on results; it admits small character-level differences such as punctuation or a one-character OCR variant and typically rejects larger omissions, such as a missing word in a short answer. We checked it by manual review of scores near the threshold (Section 3.6) and a sweep from 80 to 95 (Limitations item 5). A blank answer for a question with a gold answer is an omission (reported as a blank response in Section 4). We classify every other non-matching answer as a misassociation if it fuzzy-matches the text of any entity belonging to a different field, and otherwise as a hallucination, which is therefore a catch-all for answers not supported by any other field, including truncated or paraphrased ones. Questions linked to several answer entities are scored once, against the concatenation of all linked entities in reading order. A partial answer, such as only the first of two wrapped lines, usually falls below the threshold and is classed as a hallucination, not a misassociation, because the question's

own gold entities are excluded from the comparison (Appendix B gives worked examples). For the three graph conditions, we compute a *wrong-edge-consistency* rate, whether the pipeline's answer matches what the shown edge implies, among questions whose shown edge is itself wrong by this same fuzzy-text ruler. The question-level failure-mode buckets of Section 4.2 use the same rule; Appendix E details how this affects the deranged graph's edge-correct bucket.

Because field scores from the same document are correlated, confidence intervals come from a document-level paired bootstrap (20,000 resamples), never treating individual fields as independent. Each model-level comparison is tested once per model, so we report each family under a Bonferroni correction for its own number of tests, a per-comparison confidence level corresponding to 95% family-wise control (98.75% for a four-test family, 99.17% for six), alongside the uncorrected 95% interval.

3.6 Scorer validation

Manual review of two non-overlapping 50-item samples (one stratified across every model and condition, one drawn from scores near the fuzzy-match threshold), a spot check against the authors' own judgment rather than an independently annotated inter-rater study, agreed with the automatic verdict in all 100 cases. A small number of recurring patterns are reported as limitations rather than one-off noise (see Limitations).

4 Results

4.1 Overall effects of supplied relational context

Table 4 reports every primary contrast under our one condition-invariant prompt, each interval already Bonferroni-corrected within its four-test family. Adding the heuristic graph (F1 0.47) to flat significantly *hurts* three of four models, Claude, Qwen, and Gemini; only GPT-5's corrected interval includes zero. The deranged graph shows the same broad split, significantly hurting Claude, Qwen, and Gemini while leaving GPT-5's small negative point estimate non-significant. Gold relations, by contrast, help every model in both comparisons that isolate them: Oracle beats Flat by 4.0 to 7.9pp, and Oracle beats Heuristic by 9.1 to 27.9pp (Appendix E, Table 8), our single largest and most consistent effect, all eight corrected intervals excluding zero. We report Oracle−Flat as a secondary, post-hoc contrast rather than one of our originally locked comparisons, and read both oracle contrasts as an upper bound on faithful relay under complete, correctly labeled information, not evidence of relational reasoning in general.

Condition	Claude	GPT-5	Qwen	Gemini
Flat	75.5	79.3	77.9	79.1
Heuristic graph	59.1	76.8	58.8	61.9
Deranged graph	51.2	78.5	54.0	71.0
Oracle graph	82.0	85.9	86.7	83.1
Flat, on documents matched to:				
Heur., Oracle	75.5	79.3	78.8	79.1
Deranged	75.8	79.3	78.8	79.1
Number of documents:				
Flat	194	194	194	194
Heur., Oracle	194	194	193	194
Deranged	188	189	188	189

Table 3: Field-completion accuracy (%, fuzzy match, final answer) by condition and model, our primary evaluation. Each cell uses every document that model completed under that condition, so the document sets differ slightly (last three rows); Qwen's graph conditions cover 193 documents because one document fails to parse under every graph condition. The paired contrasts in Table 4 use only documents present in both conditions, so each equals the condition's score minus the matched Flat row, up to rounding, not the first Flat row (for Qwen, $58.8-78.8=-20.0$ and $86.7-78.8=+7.9$). Details are in Limitations item 1.

A direct interaction contrast (GPT-5's own Heuristic−Flat and Deranged−Flat delta bootstrapped against each other model's same delta on one shared resampled document set, Appendix E, Table 9) provides evidence beyond the absence of significance against zero: GPT-5 shows significantly smaller condition effects than every other model in this evaluation, in all six pairwise tests after a Bonferroni correction across the family (99.17% per-comparison CI), though the smallest (vs. Gemini on Deranged−Flat) only marginally survives.

4.2 Failure-mode decomposition

We split every scorable question into three buckets by comparing the answer implied by the heuristic graph's shown edge or edges, if any, with the gold answer under the same fuzzy rule used for scoring. For a question linked to several answer entities, the implied answer is their concatenation in reading order, so an edge set that recovers only part of the answer counts as wrong. The buckets are correct

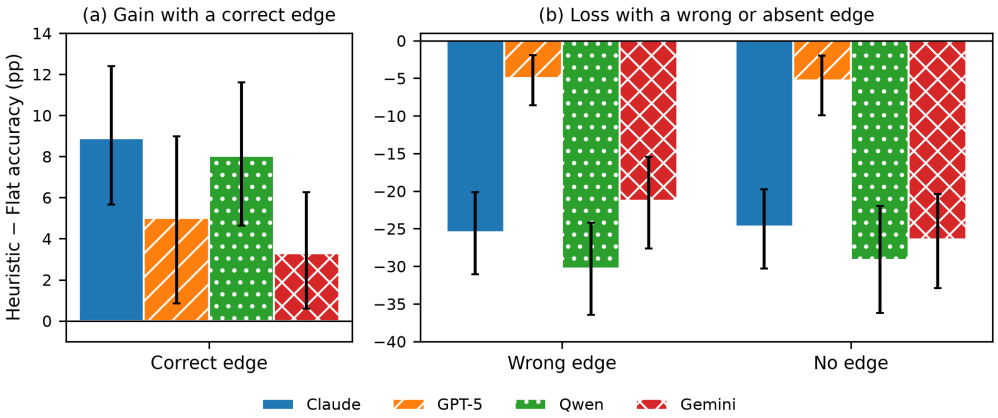


Figure 2: Heuristic−Flat accuracy difference, in percentage points, for the same questions under the two conditions, split by the heuristic relation shown. Panel (a) shows the gain where the edge is correct, and panel (b) the loss where it is wrong or absent (note the different y-axis ranges). Error bars are uncorrected 95% document-bootstrap CIs; Bonferroni-corrected significance is in the text.

Contrast	Model	Δ (pp)	Bonf. CI
Heuristic−Flat	Claude	−16.3	[−21.7, −11.4]
	GPT-5	−2.5	[−7.5, 0.7]†
	Qwen	−20.0	[−25.6, −14.2]
	Gemini	−17.1	[−22.5, −12.2]
Oracle−Flat (secondary)	Claude	+6.5	[4.4, 8.6]
	GPT-5	+6.6	[4.7, 8.5]
	Qwen	+7.9	[5.2, 11.0]
	Gemini	+4.0	[2.6, 5.6]
Deranged−Flat	Claude	−24.6	[−32.7, −17.2]
	GPT-5	−0.8	[−5.7, 2.1]†
	Qwen	−24.7	[−33.3, −16.4]
	Gemini	−8.1	[−13.3, −3.7]

Table 4: Field-completion accuracy effects and paired condition contrasts vs. Flat, each interval already Bonferroni-corrected within its four-test family (98.75% per-comparison CI). † marks the only two intervals that include zero (both GPT-5). Document counts are in Limitations item 1; Oracle−Heuristic is in Appendix E, Table 8.

(25.6–25.7% of scorable questions), wrong (32.4–32.5%), or absent (no edge shown, 41.8–42.0%, the largest bucket). Figure 2 reports the Heuristic−Flat delta within each, per model, with a document-bootstrap 95% CI. Questions with a correct heuristic edge score 3.3 (Gemini) to 8.9pp (Claude) higher under Heuristic than their matched Flat counterparts, significant at the uncorrected 95% level for all four models but surviving a Bonferroni correction across the twelve model-by-bucket tests in this family only for Claude and Qwen, not GPT-5 or Gemini. All eight wrong- and absent-edge deltas survive the same correction. Wrong- and absent-edge questions show different dominant error signatures. Wrong-edge questions show elevated cross-field misassociation, a real value copied from a different field, in 25.9% (Claude), 20.4% (Qwen), 19.4% (Gemini), and 6.3% (GPT-5) of all such questions. No-edge questions instead show elevated blank-response rates, 36.2% (Claude), 44.0% (Qwen), 40.8% (Gemini), 19.1% (GPT-5), even though the same answer-bearing entity text used by Flat remains directly present in the same representation. These percentages are shares of every question in the bucket, not just the incorrect ones. Relative to the same questions under Flat, Round 3 increases no-edge blank responses by 8.2–31.9pp and wrong-edge misassociation by 3.4–21.4pp (Appendix E, Table 11), with all eight matched penalties individually Bonferroni-significant. GPT-5's error rates stay well below the other three models in both failure buckets, consistent with its comparatively flat accuracy across every condition (Table 3); this decomposition is diagnostic rather than a controlled per-edge intervention (Limitations item 4). We exclude the deranged graph's own edge-correct bucket from analysis: 92% of it (84 of 91 cases) reflects gold-text coincidence rather than a genuine identity match (Section 3.2).

4.3 Exploratory: prompt-regime sensitivity

Our strongest sensitivity evidence compares Round 2 and Round 3, condition-invariant prompts that differ only in whether the neutral link-explanation sentence (Section 3.4) is present. The matched no-edge blank penalty and wrong-edge misassociation delta of Section 4.2 (Appendix E, Table 11)

remain positive for every model under both rounds; adding the explanation does not consistently reduce them, and for Claude both grow substantially. Round 1 was not condition-invariant and differs from Round 3 along more than one wording dimension, so we report it only as provenance, not controlled evidence: wrong-edge reliance is far lower under Round 1 for every model (Appendix D gives the figure and Table 10 the bootstrapped deltas). Because every regime was developed on this same corpus, we treat all of this as exploratory, not a causal test: the two failure signatures are present under both condition-invariant regimes examined, not shown to be prompt-invariant in general.

5 Discussion

Adding a relational channel changes what its silence means. For several models, an unlinked question is treated as unsupported even though the same entity text suffices under Flat, so absence within the channel can acquire epistemic weight it would not otherwise carry, while a wrong edge redirects the answer toward a different field's value and correct relations, unsurprisingly, still help. The elevated blank rate suggests that several models treat a missing relation as evidence against answer availability, rather than falling back to the entity text itself; a system consuming externally supplied relational context should therefore distinguish "no relation was predicted" from "no answer exists."

Susceptibility to all of this is strongly model-dependent rather than a universal LLM behavior. GPT-5 shows little heuristic or deranged degradation, the lowest wrong-edge-consistency rate by a wide margin, and error rates well below the other three models in both failure buckets, while Claude, Qwen, and Gemini are all substantially affected. We report this heterogeneity as a result in its own right rather than speculate about its cause.

This also bears on anyone building a pipeline on an automatic relation extractor. A linker with F1 around 0.47 is not automatically safe to add, and its downstream effect also appears sensitive to prompt regime in our exploratory comparison. A single run cannot distinguish a model property from an artifact of one prompt's wording, which is why Appendix A gives the full prompt history.

6 Conclusion

Imperfect relational context can alter downstream LLM extraction in ways that aggregate accuracy obscures. In our FUNSD evaluation, wrong and missing heuristic relations are associated with distinct error signatures whose magnitudes vary across models and remain visible under both condition-invariant prompt regimes examined. Future work should manipulate relation correctness and coverage independently, repeat generations, and test additional datasets.

Limitations

1. **Document counts.** Table 3 gives the documents behind each cell: 194 for Flat, Heuristic and Oracle (193 for Qwen3.6-27B, one persistent parse failure under every graph condition), and 188 or 189 for Deranged, which also excludes five documents that cannot be deranged (Section 3.2) and, for Claude, one further parse failure. Paired contrasts in Table 4 use the intersection of the two conditions' documents, so each equals the difference of the matched Flat rows, up to rounding. Scoring every persistent-parse-failure cell as entirely wrong instead of excluding it shifts every affected Table 4 headline by at most 0.5pp (Qwen3.6-27B's Oracle−Flat, whose excluded document has 41 scorable questions against a corpus average of 14) and every other by at most 0.1pp.

2. **Adaptive prompt development, single generation, and provider versions.** The final prompt (Section 3.4) was informed by this corpus's own results from two preliminary rounds, so the primary round is a frozen-before-execution evaluation, not a hypothesis specified before any data existed. Every cell in every round is one generation, so the document-level bootstrap captures variability across documents, not across repeated generations, and we treat low-single-digit effects cautiously even when their interval excludes zero (GPT-5's heuristic and deranged estimates did not survive even the uncorrected test). None of the four model identifiers is a dated, pinned snapshot (Appendix A), which affects only the exploratory cross-round comparison (Section 4.3). The three prompt regimes also differ in more than one wording dimension, so that comparison is exploratory, not causal.

3. **Gold upstream information, not an end-to-end pipeline.** Every condition receives

FUNSD's gold entity boundaries, type labels, and bounding boxes, which isolates relational structure as the one manipulated variable but means none of our numbers speak to a pipeline that also solves segmentation, OCR, or typing. The heuristic graph is a layout-proximity rule over those gold entities, not an end-to-end relation extractor, and a stronger or weaker linker could shift its result either direction. We also do not compare against feeding the page image directly or against HTML or Markdown produced by layout analysis, so we cannot say whether these failure modes would appear, shrink, or grow under those inputs.

4. **Detectability and plausibility are diagnosed, not independently manipulated.** Heuristic and deranged conditions differ in spatial plausibility, semantic compatibility, and structure at once (equivalently, coverage and correctness: ≈58% coverage at F1 ≈0.47 versus 100% coverage at near-zero correctness), so their contrast does not isolate any one factor. A degree-preserving derangement keeps each entity's participation count, multi-value cardinality, and graph topology, so it is not information-free noise; the few identity-correct and text-duplicate edges it can produce are disclosed and excluded from any wrong-edge-only diagnostic (Section 3.2). Relation-status buckets (Section 4.2) are diagnostic, since status is observed within the heuristic representation and not manipulated per question.

5. **Additional design limitations.** We ran no extract-then-re-extract condition separating a second call's own effect from a second read over relational structure, so our correction-rate pattern is descriptive, not evidence that the relational representation drives verification behavior. We merged our original development/test split into one set partway through, so every document was eligible for both development and the final numbers, weaker than an untouched test set; the heuristic has no learned parameters, though its design was developed within the same study. Hand validation also showed the scorer cannot distinguish a correct but reordered multi-value answer from a wrong one; an order-insensitive rescoring and a fuzzy-threshold sweep (80–95) each change Table 4's Heuristic−Flat and Deranged−Flat deltas, the two contrasts we checked this way, by at most 1.6pp.

6. **Scope of generalization.** All results are on one English-language, form-specific benchmark, FUNSD, with one serialization of each representation, one model roster, and one pipeline architecture; a different choice along any of these axes could change the magnitude or direction of any effect reported here. We use the original FUNSD annotations, whose known block-level limitations are addressed by EC-FUNSD (Zhang et al., 2026), so replication on the corrected annotation set is a natural next step.

Ethical Considerations

Data provenance and use. FUNSD's source documents come from the Truth Tobacco Industry Documents archive, made public through litigation disclosure decades ago, and some forms carry real names, signatures, and dates from that era. We collected no new data. FUNSD is released under a noncommercial research and education license, and we send only document text, never images, to four commercial LLM APIs, for noncommercial academic research. **Human involvement.** There was no new human-subject data collection or crowdsourced annotation; involvement was limited to the authors' review of a sample of scored outputs (Section 3.6). **Reproducibility.** Code, the Appendix A.2 prompt templates, and the aggregate analysis files behind every table are released at the repository linked in the Introduction under an open-source license. FUNSD is not redistributed apart from one small sample annotation used as a test fixture (a script downloads the dataset from its official source), and per-question model outputs are withheld because they contain FUNSD text, so reproducing raw runs requires API access to the four models.

Declaration on Generative AI

Portions of this manuscript were drafted and revised with writing assistance from Claude (Anthropic), which was also used for coding assistance during the experimental pipeline's implementation. All experimental design, analysis, results, and conclusions are the authors' own, reviewed and verified before submission.

References

Tek Raj Chhetri, Yibei Chen, Puja Trivedi, Dorota Jarecka, Saif Haobsh, Patrick Ray, Lydia Ng, and Satrajit S. Ghosh. 2025. STRUCTSENSE: A task-agnostic agentic framework for structured information extraction with human-in-the-loop evaluation and benchmarking. *arXiv preprint arXiv:2507.03674*.

Gaye Colakoglu, Gürkan Solmaz, and Jonathan Fürst. 2025. Problem solved? information extraction design space for layout-rich documents using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Xinbang Dai, Yuncheng Hua, Tongtong Wu, Yang Sheng, Qiu Ji, and Guilin Qi. 2025. Large language models can better understand knowledge graphs than we thought. *Knowledge-Based Systems*, 312(3):113060.

Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. 2022. Doc2Graph: A task agnostic document understanding framework based on graph neural networks. In *ECCV Workshops*.

Diego Gosmar and Giovanni Zenezini. 2026. MADP: A multi-agent pipeline for sustainable document processing with human-in-the-loop. *arXiv preprint arXiv:2605.17159*.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Yangyang Hui, Jiahao Liu, Qi Zhang, Tianyi Zhou, and Yonghong Song. 2025. SEG-Doc: A simple yet efficient graph neural network framework for document key information extraction. *Neurocomputing*, 647.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR Workshop on Open Services and Tools for Document Analysis (ICDAR-OST)*.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. FormNet: Structural encoding beyond sequential modeling in form document information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pritika Ramu, Sijia Wang, Lalla Mouatadid, Joy Rimchala, and Lifu Huang. 2024. RE^2: Region-aware relation extraction from visually rich documents. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yuanfu Sun, Zhengnan Ma, Yi Fang, Jing Ma, and Qiaoyu Tan. 2025. GraphICL: Unlocking graph learning potential in LLMs through structured prompt design. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2440–2459.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hieu M. Vu and Diep Thi-Ngoc Nguyen. 2020. Revising FUNSD dataset for key-value detection in document images. *arXiv preprint arXiv:2010.05322*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chong Zhang, Yixi Zhao, Yulu Xie, Chenshu Yuan, Yi Tu, Ya Guo, Mingxu Chai, Ziyu Shen, Yue Zhang, and Qi Zhang. 2026. Unveiling the deficiencies of pre-trained text-and-layout models in real-world visually-rich document information extraction. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1–16, Rabat, Morocco. Association for Computational Linguistics.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2023. Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia*, 25:6743–6755.

A Full prompt provenance

We ran this ablation three times on the same 194-document corpus. **Round 1**'s graph prompt explained what a link means, instructed treating it as a primary guide, and warned that shown links may be wrong and should not be followed blindly. This text was absent from the flat and raw prompts, so it confounded structure with an explicit skepticism instruction. **Round 2** removed that instruction but, a later diff showed, also removed the generic link explanation along with it, so Round 1 and Round 2 differ along more than one dimension. Under Round 2's condition-blind but under-explained prompt, Qwen3.6-27B and Gemini 3.7 Flash showed a severe blank-refusal pattern specifically on questions whose heuristic graph offered no

edge, consistent with treating an unexplained absence of a links field as license to answer null rather than falling back to the flat entity text present in the same representation. Restoring the explanation in Round 3 did not reliably reduce this pattern (Section 4.3). **Round 3** (reported as primary throughout the main text) restores one short, neutral sentence, identical across all conditions, explaining that the representation may include structural information such as span labels, positions, or links, that a link when present associates a question with an answer, and that some questions have none. Its generic first half is not data-informed; the clause about questions with no link directly responds to Round 2's failure on this same corpus, a distinction we disclose rather than calling the whole sentence either fully generic or fully post hoc. A banned-word list for the input-specific paragraph (correct, incorrect, reliable, unreliable, candidate, trust, distrust, primary guide, plausible, blindly, fall back, and inflections), a check against gating any instruction on link presence, and a regression test that the assembled extract and verify prompts are byte-identical across all five conditions once the document representation is removed all carry over from Round 2. The word list does not apply to the verify prompt's own instructions, which legitimately use words such as "correct". A manifest audit (`git diff` across every module and the dataset file between the two later rounds' launch commits) confirms the prompt files are the only source difference. We describe our analysis plan as frozen before Round 3's execution and informed by Rounds 1 and 2 on this same corpus, not pre-registered in the sense of specified before any data existed.

Scope split. The raw-versus-flat background comparison (Appendix C) is reported from Round 2 only, the first condition-invariant version of that comparison; Raw shares no structural field with the link-definition sentence, so it is unaffected by Round 2's no-edge problem, but it was not rerun under Round 3's exact wording. Flat's own accuracy is stable within 0.6pp for every model between Round 2 and Round 3, which we use as a measured proxy for Raw's likely stability, not a direct measurement of it. All of Section 4 and Section 4.2 come from Round 3 alone; no table merges Round 2 and Round 3 columns as directly comparable.

Model versions across rounds. All three rounds used the same model identifiers (Section 3.3) across roughly a four-day window; a manifest audit confirms our own code, data, and requested identifiers were unchanged. None of the four identifiers is a dated, pinned snapshot, so we cannot independently verify a provider's served implementation was itself unchanged across our calls.

A.1 Model and API reproducibility

Table 5 lists the exact model identifiers, reasoning settings, and call parameters.

A.2 Literal prompt templates

Every extract prompt shares this fixed scaffold across all three rounds and all five conditions, substituting only the emphasized input-specific paragraph below and the document itself:

> "You are a document-processing assistant completing a structured form-filling task.
>
> *[input-specific paragraph, below]*
>
> Instructions: Answer using ONLY information present in the document below. Do not infer, guess, or fabricate any value that is not explicitly present. If you cannot find an answer for a question in the document, output an empty string `""` for that answer, do not leave it out of the list entirely. Respond with ONLY a JSON array, no other text, in this exact format: `[{"question": "<question text as it appears>", "answer": "<answer text, or empty string>"}, ...]`
>
> Document: [the rendered representation]"

The verify step re-reads the same input-specific paragraph alongside the model's own previous answer, under the same JSON-array output format. Its full template, identical across rounds and conditions apart from the same input-specific paragraph, is:

> "You are re-checking your own previous answer to a document form-filling task, to catch and correct any mistakes before finalizing.
>
> *[input-specific paragraph]*
>
> Your previous answer was: *[the model's JSON answer from the extract step]*
>
> Re-read the document below and your previous answer above. For each question, check whether the answer is actually supported by the document, correct any answer that is wrong, ungrounded, or inconsistent, and leave correct answers unchanged.
>
> Instructions: Base corrections ONLY on information present in the document below. Do not infer, guess, or fabricate. If an answer should be blank because the document doesn't contain it, output an empty string `""`. Respond with ONLY a JSON array, no other text, in this exact format: `[{"question": "<question text as it appears>", "answer": "<answer text, or empty string>"}, ...]`
>
> Document: [the rendered representation]"

Model	Provider	Requested ID	Reasoning setting
Claude Sonnet 5	Anthropic	`claude-sonnet-5`	`thinking`: adaptive, `effort`: low
GPT-5	OpenAI	`gpt-5`	`reasoning_effort`: low
Qwen3.6-27B	Groq	`qwen/qwen3.6-27b`	`reasoning_effort`: default
Gemini 3.7 Flash	Google	`gemini-3.7-flash`	`thinking_level`: low

Table 5: Exact requested model identifiers and reasoning settings, held constant across all conditions and rounds. Temperature: Claude and GPT-5 reject any explicit value (omitted); Qwen and Gemini use 0.0. Max output tokens: Claude 128,000 (provider ceiling, required explicitly by the API); GPT-5 and Gemini use the provider default (uncapped); Qwen is capped at 16,384, Groq's own hard per-call maximum for this model. All Round 3 (primary) calls were made on 2026-09-03.

Round 1 input-specific paragraph, flat and raw conditions (Round 1 only; Rounds 2 and 3 use one paragraph for all conditions, below). Flat: "Below is a JSON list of labeled text spans extracted from a scanned form. Each span has a 'text' value and a 'label' of 'question' (a field name), 'answer' (a field value), 'header' (a section title), or 'other'. The spans are NOT linked to each other, you must match each 'question' span to the 'answer' span you judge most likely to be its value, based on the text content, position (each span includes its 'box' pixel coordinates), and typical form layout conventions." Raw: "Below is the plain text content of a scanned form, listed in reading order (top-to-bottom, left-to-right). It has no field labels or structure markup, you must recognize which parts of the text are field labels ('questions') and which are their corresponding values ('answers') yourself, based on the wording and layout implied by the reading order."

Round 1 input-specific paragraph, graph conditions. (Rounds 1 and 2 labeled the Heuristic and Deranged conditions "Predicted" and "Shuffled".) "Below is a JSON structure of labeled text spans extracted from a scanned form, plus a 'links' list of [id, id] pairs indicating which spans the document's structure connects (typically a question span to the answer span that follows it). Each span has a 'text' value and a 'label' of 'question', 'answer', 'header', or 'other'. Use the given links as your primary guide to match each 'question' span to its answer, but note that this linking structure was produced automatically and may contain errors, if a linked answer's text doesn't plausibly correspond to its question, use your own judgment about the document's content instead of following the link blindly."

Round 2 input-specific paragraph (all five conditions, condition-blind but under-explained; Appendix A discusses its no-edge failure). "Below is a machine-readable representation of one scanned form. Use it to identify each field name (a 'question') and the value that answers it (an 'answer')."

Round 3 input-specific paragraph (all five conditions, primary throughout the main text). "Below is a machine-readable representation of one scanned form. The representation may include structural information such as span labels, positions, or links between spans, depending on how the document is represented. A link, when present, associates a question span with an answer span. Some question spans do not have an associated link. For each field name (a 'question'), determine the value that answers it (an 'answer') using all information available in the representation below."

B Scoring examples

Table 6 shows how the scorer of Section 3.5 treats typical answers, using the constructed form of Table 2 extended with a two-line answer. Two examples need a note. A multi-entity answer is scored as one string, so giving only the first line is incorrect and, since that line belongs to the question's own gold answer, is not counted as a misassociation. A misassociation requires that the answer match an entity of a *different* field, so copying "Acme" for Name is a misassociation but inventing "Acme Corp." for Name is a hallucination.

C Background: Raw Text versus Flat Structuring

Moving from raw text to flat, typed, laid-out entities produces a large gap under Round 2's condition-invariant prompt, 14.8 (Gemini) to 23.0 (Claude) points, all Bonferroni-significant across the four models. This is a representation comparison, changing segmentation, labels, layout, and serialization together, not an isolated test of "semantic structure" alone.

Question	Gold answer	Model answer	Score
Name	J. Smith	j. smith	correct (exact after normalization)
Name	J. Smith	J Smith	correct (fuzzy, ratio ≥ 85)
Name	J. Smith	(blank)	omission
Name	J. Smith	Acme	misassociation (matches Brand's answer)
Name	J. Smith	Acme Corp.	hallucination (matches no field)
Address	12 Main St. Ohio	12 Main St.	incorrect, hallucination (partial; own entity excluded)
Address	12 Main St. Ohio	12 Main St. Ohio	correct (both lines)

Table 6: Scoring examples on constructed answers. Accuracy counts only correct rows.

Condition	Claude	GPT-5	Qwen	Gemini
Raw	53.1	57.5	55.8	64.1
Flat	76.1	79.7	77.9	79.0

Table 7: Round 2 background comparison, not part of the primary Round 3 evaluation (Appendix A).

D Exploratory prompt-regime sensitivity, full figure

E Additional statistical analyses

E.1 Dataset and representation details

The heuristic's 35-document validation-subset precision and recall (0.53, 0.43) come from 305 true positives of 576 predicted and 716 gold edges. The degree-preserving derangement can occasionally map an edge back to a technically correct target (7 of 2,609 edges) or to a different entity whose text happens to duplicate the gold answer's text, most often a repeated date or form number (84 of 2,609); both are disclosed and excluded from any diagnostic that requires a genuinely wrong, text-distinguishable edge (Section 3.2).

Model	Δ (pp)	Bonf. CI
Claude	$+22.8$	$[17.9, 28.1]$
GPT-5	$+9.1$	$[5.9, 13.5]$
Qwen	$+27.9$	$[22.6, 33.2]$
Gemini	$+21.1$	$[15.9, 26.8]$

Table 8: Oracle−Heuristic, the same intersection-of-scorable-documents contrast as Table 4, four-test family (98.75% per-comparison CI). All four tests exclude zero.

Contrast	vs.	Δ (pp)	Bonf. CI
Heuristic−Flat	Claude	$+13.8$	$[5.6, 20.7]$
	Qwen	$+17.6$	$[11.4, 23.7]$
	Gemini	$+14.6$	$[7.1, 21.4]$
Deranged−Flat	Claude	$+23.8$	$[13.6, 33.6]$
	Qwen	$+24.0$	$[13.8, 34.0]$
	Gemini	$+7.3$	$[0.1, 13.9]$

Table 9: GPT-5 interaction contrast: GPT-5's own (contrast−Flat) delta minus each other model's same delta, six-test family (99.17% per-comparison CI). Positive values mean GPT-5 is less affected than the comparison model.

Condition	Model	R1	R3	Δ (Bonf. CI)
Deranged	Claude	0.8%	28.2%	$+27.4$ $[19.7, 35.8]$
	GPT-5	0.1%	1.2%	$+1.1$ $[0.2, 2.3]$
	Qwen	0.8%	31.1%	$+30.3$ $[21.5, 39.0]$
	Gemini	0.1%	11.0%	$+11.0$ $[6.2, 16.3]$
Heuristic	Claude	4.8%	25.7%	$+21.0$ $[14.8, 27.8]$
	GPT-5	3.2%	7.0%	$+3.9$ $[1.6, 6.6]$
	Qwen	5.0%	30.8%	$+25.8$ $[18.5, 33.3]$
	Gemini	2.0%	21.6%	$+19.6$ $[12.9, 27.0]$

Table 10: Wrong-edge-consistency rate, Round 1 (R1) vs. Round 3 (R3), restricted to documents scored in both rounds, four-model family (Bonf. CI). Every model's increase excludes zero.

Metric	Model	Round 2	Round 3	Δ R3−R2 (Bonf. CI)
No-edge blank penalty	Claude	$+4.6$	$+24.0$	$+19.5$ $[13.2, 26.5]$
	GPT-5	$+4.8$	$+8.2$	$+3.4$ $[-2.2, 10.3]$
	Qwen	$+41.3$	$+31.9$	-8.7 $[-17.6, 0.0]$
	Gemini	$+35.2$	$+30.1$	-5.1 $[-15.2, 4.3]$
Misassociation delta	Claude	$+8.6$	$+21.4$	$+12.8$ $[7.5, 18.3]$
	GPT-5	$+2.2$	$+3.4$	$+1.2$ $[-1.1, 3.8]$
	Qwen	$+23.9$	$+17.9$	-5.9 $[-12.6, 0.4]$
	Gemini	$+15.4$	$+16.2$	$+0.8$ $[-5.3, 7.0]$

Table 11: No-edge blank penalty and edge-wrong misassociation delta (both, Heuristic minus Flat, matched on the relevant bucket's questions), Round 2 vs. Round 3, and their bootstrapped Round3−Round2 change (four-model family, Bonf. CI). Round 2 and Round 3 values are each individually Bonferroni-significant against zero for every model (CIs not shown here).

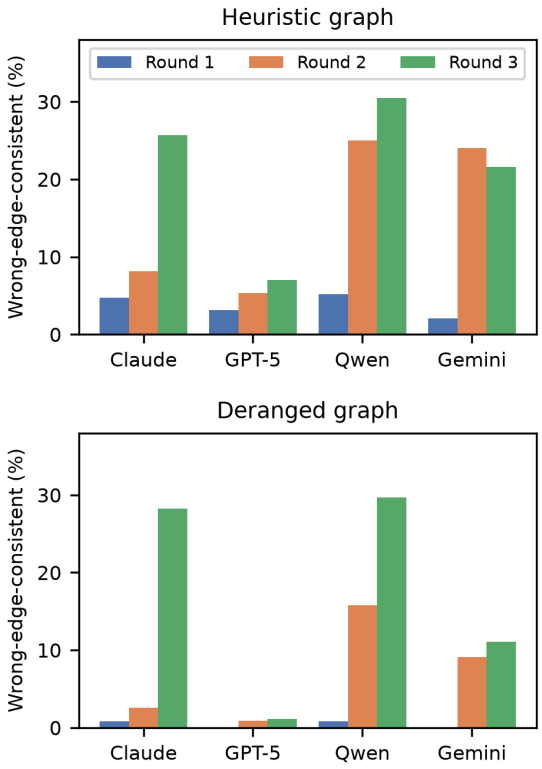


Figure 3: Wrong-edge-consistency rate (among applicable wrong-edge questions) across Rounds 1, 2, and 3, by model, for the heuristic and deranged graph conditions separately. Grouped bars, not a continuous line, since the three prompt regimes are qualitatively distinct rewrites, not points on a single dose axis. See Section 4.3 for discussion.