

What 1,114 Documents Reveal About Schema-Driven Extraction: Models, Pipelines, and the Real-vs-Synthetic Gap

Frank Thomas

Koji Technologies Inc

frankthomas.lg@gmail.com

Andrei Vince

University of North Carolina at Charlotte

andreivince21@gmail.com

Abstract

Extracting structured data from documents is a core task in finance, insurance, legal, and healthcare, yet empirical studies comparing extraction approaches across domains remain scarce. We assemble a corpus of 1,114 documents across 12 categories spanning 8 domains, with 7,197 fields annotated with ground-truth values, and use it to study how extraction accuracy varies across models, pipeline architectures, and document sources. We evaluate with a type-aware methodology that scores every field as a correct extraction, a miss, a hallucination, or a correct absence, so that declining to extract is distinguished from extracting wrongly. Our central finding is that aggregate accuracy conceals critical divergence along every axis: synthetic documents overestimate accuracy by 38–45pp on insurance certificates ($n=18$ real documents); three models score within 1pp overall but diverge by 31pp per category; and two pipeline architectures diverge by 26pp on contracts—with the advantage not predicted by document length. An independently built extraction system, run under a protocol fixed in advance, agrees with our ground truth on 82–87% of fields across three models spanning two providers, providing evidence that the corpus is not calibrated to a single implementation. The corpus and evaluation tooling are released to enable reproducible comparison.

1 Introduction

Commercial document-extraction products routinely advertise accuracy above 95%: 99% field-level accuracy across 250+ document types (Docsumo, 2026), accuracy rates of 99.5% (HyperScience, 2026), and invoice extraction accuracy up to 99% (Doxis AI Solutions, 2026) are all current product-page claims. None of the three publishes the measurement protocol behind its figure. This makes such claims difficult to compare or falsify: a single “95%” may denote character accuracy,

field-level accuracy, or the share of documents extracted with zero errors—quantities that diverge substantially on real documents. A customer cannot determine which is meant, and practitioners evaluating or building these systems lack empirical evidence about how accuracy actually varies across document types, models, and pipeline architectures.

Existing evaluation resources are either small—ExtractBench (Ferguson et al., 2026) covers 35 documents across 5 schemas and 4 domains—or single-domain with synthetic ground truth, as with VAREX (Barzelay et al., 2026) and its 1,777 government forms. Neither supports cross-domain analysis. We therefore assemble a larger and more diverse corpus and use it as an instrument: the corpus is the means, and the object of study is how extraction accuracy behaves across models, pipeline architectures, and document sources.

We therefore ask four questions about schema-driven document extraction, each concerning *field-level extraction accuracy*:

1. Does a single aggregate accuracy figure adequately summarize a system, or does accuracy diverge across document categories?
2. How much does the choice of extraction model affect accuracy, and is the most accurate model the same across categories?
3. How much does pipeline architecture affect accuracy, and what predicts which architecture wins?
4. Do synthetic documents predict accuracy on real documents of the same type?

Our contributions: (1) A **cross-domain corpus** of 1,114 documents across 12 categories and 8 domains, with 7,197 field-level annotations. (2) A **type-aware evaluation methodology** that scores every field as a correct extraction, a miss, a hallucination, or a correct absence (§4), separating failures to extract from correct abstention. (3) **Empirical findings** from two pipelines across six

models spanning five families: a 38–45pp real-vs-synthetic gap, 31pp per-category model divergence, and pipeline gains that reverse by category.

2 Related Work

Document understanding. DocVQA (Mathew et al., 2021), FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), and DUDE (Van Landeghem et al., 2023) evaluate document understanding but do not target schema-driven extraction. Extract-Bench (Ferguson et al., 2026) (35 docs, 5 schemas, 4 domains) is too small for per-domain analysis. VAREX (Barzelay et al., 2026) (1,777 docs) is single-domain with synthetic GT.

Document understanding models. LayoutLM (Xu et al., 2020) and LayoutLMv2 (Xu et al., 2021) pre-train on document layout for visual understanding (bounding box localization, form field recognition). Our corpus targets a complementary axis: given parsed text and a typed schema, how accurately can a system populate every field? CORD (Park et al., 2019) provides 11K receipts for post-OCR extraction; DocILE (Šimsa et al., 2023) offers 6.7K business documents for key information localization. Both are single-domain; our corpus spans 12 categories.

Extraction pipelines. Modern pipelines partition documents and route fields to relevant sections (Lewis et al., 2020; Sarthi et al., 2024). Self-RAG (Asai et al., 2024) demonstrated that retrieval filtering improves generation by avoiding irrelevant context. SLIDE (Singh et al., 2025) uses sliding windows; DocAgent (Sun et al., 2025) employs multi-agent querying; PARSE (Shrimal et al., 2025) optimizes the schema itself. These have been evaluated on task-specific datasets; our corpus enables controlled cross-domain comparison.

Evaluation methodology. Field-level evaluation has a long history (Tjong Kim Sang and De Meulder, 2003). Our methodology extends standard approaches with type-aware comparison and a null-aware scoring rule that distinguishes extraction failures from correct absence.

3 Corpus Design

The corpus contains 1,114 documents across 12 categories spanning 8 domains (Table 1); per-category field counts and the pinned corpus version are given in Appendix B.

Category	Docs	Real	Synth	Domain
SEC filings	288	288	0	Finance
Invoices	155	5	150	Retail
Ins. claims	152	22	130	Insurance
Contracts	100	100	0	Legal
Medical records	100	100	0	Healthcare
Ins. policies	97	17	80	Insurance
Legal filings	61	61	0	Legal
Ins. certificates	61	21	40	Insurance
Receipts	52	52	0	Retail
IRS forms	20	0	20	Tax
Adversarial	25	0	25	Adversarial
Multi-format	3	3	0	Multi-format
Total	1,114	669	445	8 domains

Table 1: Corpus composition. “Ins.” = Insurance. Multi-format ($n=3$) is excluded from results tables due to insufficient sample size. Three certificates listed here as real are reclassified as synthetic by manifest provenance (§5.3); the evaluable split is 18/43 for certificates and 666/448 corpus-wide, and that is the split used in all real-vs-synthetic analyses.

Real documents. The 669 real documents come from six public sources. SEC filings (288) are EDGAR 10-K, 10-Q, 8-K, DEF 14A, S-1, 20-F and 6-K submissions, and contracts (100) are material contracts filed as EDGAR 8-K Exhibit 10. Both are mandatory public disclosures; they are authored by the filing registrant rather than the SEC, so they are not U.S. Government works and carry no explicit licence. Medical records (100) are MTSamples clinical transcriptions spanning 15 specialties, shared under the source’s educational-use terms (see Ethics Statement). Legal filings (61) are court opinions from CourtListener and the Caselaw Access Project, receipts (52) are SROIE scanned receipts under CC BY 4.0, and 60 insurance documents—certificates, claims and policies—come from government and state-agency sites. The remaining 8 are 5 invoices and 3 multi-format documents from the same government sources. Retrieval URLs and per-source licence terms are in Appendix A.

Document structure. Each document consists of three components: (1) **parsed content**—pre-parsed markdown text, isolating extraction accuracy from parsing accuracy (a deliberate design choice: OCR quality affects SROIE receipts but is not the variable under test); (2) **schema**—a YAML file defining field names, types (string, number, date, array, enum), descriptions, and optional extraction hints (category labels, keyword patterns, positional preferences); (3) **expected output**—

JSON with ground-truth values for every annotated field. Fields with unknown ground truth are omitted rather than set to null, so the evaluation does not penalize correct extractions that were not annotated. This means the denominator for each category reflects only fields with verified ground truth, not the full schema width.

Synthetic documents. 445 synthetic documents target failure modes that real documents do not cover densely: array extraction edge cases (invoices), state-specific form variation (insurance claims), multi-policy nested structures (insurance certificates), and adversarial inputs (blank documents, OCR noise, wrong-schema application, prompt injection, non-English text, mixed-language inputs).

Ground truth. For authoritative-index sources (EDGAR, SROIE), GT derives from the source index. For other categories, GT is seeded by GPT-4o-mini extraction and reviewed by a single annotator per category. Across the 1,100-document subset for which the comparison was run (7,121 fields), 829 (11.6%) diverge between mini’s output and reviewed GT; these comprise 349 annotator enrichments (mini returned null, annotator added value), 245 value disagreements, 152 hallucination rejections, and 83 array mismatches. The 14 adversarial documents outside that subset are not covered by it.

Independent validation. To assess GT reliability, a second author built an independent extraction pipeline without access to our implementation, prompts, or results. After a blind 10% pilot ($n=110$, 89.1% concordance), the same locked protocol—fixed before the pilot—was run on the 1,100-document snapshot (7,121 fields), achieving 86.6% observed field agreement with GT using GPT-4o-mini. Of the 13.4% disagreements: 7.5% are value mismatches, 3.4% are hallucinations by the independent system, and 2.4% are missing fields—consistent with extraction difficulty rather than GT error. Per-category concordance ranges from 68.8% (contracts) to 100% (IRS forms, multi-format). The lower contracts concordance (68.8%) is consistent with our cross-pipeline finding that full-document extraction scores 55.6% on these long-form contracts (§5.5), reflecting extraction difficulty rather than GT unreliability. Two further models score 82.1% (GPT-5.4-nano) and 86.1% (MiniMax M3) on the same snapshot—

three models spanning two providers, two with no role in GT creation, converging at 82–87%. The model both architecturally and organizationally independent of the seed (MiniMax M3, 86.1%) lands within 0.5pp of the seeding model (86.6%), putting any shared-model advantage well under a point. We report this as *locked-protocol independent extraction concordance*, not human-human IAA.

Blind second-annotator agreement. A second annotator, blind to ground truth, re-annotated a stratified sample of 60 real documents. Across 371 fields, agreement with GT is **81.9%**, ordered by schema ambiguity rather than document difficulty (SEC filings 97.6% to legal filings 69.1%). Counting correct absences as agreements inflates this: 85 of the 304 agreements are cases where both ground truth and annotator recorded an absence. On fields where either party asserted a value, agreement is 76.6%. Cohen’s κ on enumerated fields averages 0.59, but `governing_law` scores 0.00. That figure is a canonicalisation artifact, not disagreement: the annotator wrote “State of X” where ground truth reads “X” in all twelve cases, identifying the same jurisdiction every time. The same surface mismatch inflates the extraction failures §5.5 reports for that field. Appendix F gives the per-category breakdown.

These probes are human verification of ground truth, not two-annotator agreement measured from scratch; corpus-scale multi-annotator IAA remains future work. Adjudication and spot-check detail is in Appendix F.

4 Evaluation Methodology

We measure **field-level accuracy**: correct fields divided by total annotated fields, reported as both micro-average (weighted by category size) and macro-average (unweighted mean). Extracted values are compared with type-specific normalization: dates to YYYY-MM-DD, numbers with tolerance 0.01, strings with optional Levenshtein similarity (0.85 threshold), and arrays order-independently.

Four-way null semantics. We score each field into exactly one of four outcomes: **correct extraction** (a verified value, correctly extracted), **missing** (a verified value the system failed to return; false negative), **hallucinated** (a value returned where ground truth is absent; false positive), and

correctly absent (ground truth is absent and the system correctly abstained; true negative). Only correct extractions and correct absences count as accurate. We call this *four-way null semantics*. The term denotes a null-aware taxonomy over *extraction outcomes*; it should not be confused with four-valued null logic in database theory (Codd, 1990), which distinguishes kinds of missing *value* (applicable vs. inapplicable) rather than kinds of extraction outcome.

Because correct absences count as accurate, categories with many absent fields can score higher than their extraction-of-present-values performance alone would suggest. We therefore also report **present-fields-only accuracy**, restricted to fields with a non-null ground-truth value (Appendix D); the two metrics differ by 0.3pp corpus-wide but by up to 7.3pp within a category.

As a robustness check, excluding the adversarial category yields 88.3% micro-average (vs. 88.4% with it), indicating negligible inflation. Across 3 runs (GPT-4o-mini, temperature 0), variance (half-range across 3 runs, not a confidence interval) is ± 0.05 pp corpus-level and ≤ 0.55 pp per category.

5 Experiments

5.1 Extraction Pipeline

We evaluate a schema-driven pipeline: parse $\rightarrow$ route (score sections per-field using *schema hints*—the schema’s keyword and positional annotations, applied to section selection rather than to the extraction prompt—and select the top-scoring sections) $\rightarrow$ extract (LLM with routed sections) $\rightarrow$ validate. We refer to this architecture as **chunk-route-extract**: the parsed document is split into sections (chunks), each schema field is routed to the sections predicted to contain it, and extraction runs over only those routed sections rather than the whole document. Its counterpart, the **full-document** baseline (§5.5), sends the entire document to the model in a single call. Full configuration details are in Appendix H.

A worked example. `sroie_000` is a scanned receipt that parses into four sections (merchant, receipt, items, totals). Its schema carries ground truth for three fields. Routing selects the merchant, receipt and totals sections; extraction over those returns `date=2018-12-25`, matching after normalisation from 25/12/2018, and

Category	Docs	Accuracy
IRS forms	20	100.0%
Ins. policies	97	97.9%
Adversarial	25	93.3%
Ins. certificates	61	90.8%
SEC filings	288	88.6%
Ins. claims	152	88.1%
Invoices	155	87.9%
Medical records	100	86.2%
Legal filings	61	85.5%
Contracts	100	81.6%
Receipts	52	57.8%
Micro / Macro	1,111	88.4% / 87.1%

Table 2: Baseline accuracy (GPT-4o-mini, no techniques). 42-point spread from receipts to IRS forms.

`total_amount=9.0`, matching within tolerance. For `merchant_name` it returns null where ground truth reads `BOOK TA .K (TAMAN DAYA) SDN BHD`, a thermal-print header the OCR renders inconsistently; under four-way null semantics that is a *miss*, not a correct absence, so the document scores 2/3. The trace shows both behaviours the paper measures at scale: normalisation absorbing format variation, and abstention on a degraded field charged as an error. Four techniques may be enabled on top of this architecture—context chunks, extraction hints, gap-fill retry and wave extraction—each defined in Appendix H.

5.2 Single-Model Baseline

The 42-point spread reflects fundamental differences in extraction difficulty. IRS forms have rigid structure and unambiguous field positions. Scanned receipts contain thermal-print OCR artifacts, skewed layouts, and minimal structural cues—accuracy is limited by source scan quality. The adversarial category (93.3%) scores counterintuitively high because most adversarial documents test graceful null-return, which counts as correct under four-way null semantics.

5.3 Real vs. Synthetic Gap

Insurance certificates show a -44.5 pp gap: real documents score 53.2% (± 13.7 , $n=18$ documents / 47 annotated values) vs. synthetic 97.7% (± 2.0 , $n=43$ documents / 258 values). Table 1 lists 21 real certificates, but 3 are reclassified as synthetic by manifest provenance, leaving 18 evaluable real documents. Per-field analysis reveals two layout failures: `certificate_holder_name` (41.7% vs. 100%—real ACORD forms place the

holder in an unlabeled bottom-corner box) and `policies` array (44.4% vs. 88.4%—real forms use merged-cell grid tables). The aggregate gap across all evaluable documents is 6.7pp (not stratified by category, so it partly reflects differing category difficulty) (real 85.2% $\pm$ 1.3 vs. synthetic 91.9% $\pm$ 1.2, non-overlapping 95% Wilson CIs). The 95% Wilson interval on the real arm is 39.2–66.7%. The 47 are not a uniform sample of the schema; standardising to the synthetic field mix reduces the gap to **38.2pp** (Limitations). The two other categories holding both arms show the same direction at smaller magnitude—`policies` +17.5pp, `claims` +17.0pp (Appendix G); the gap is large and directionally clear, but field-level Wilson intervals may overstate confidence because fields within the same document are correlated.

The mechanism matters more than the magnitude. Our synthetic certificates are generated from schema-consistent templates, and both failing fields fail for the same reason: the templates reproduce the *content* of an ACORD certificate but not its *layout variability*—unlabeled positional boxes and merged-cell grid tables that real forms use and templates do not. The gap is therefore not evidence that synthetic data is unusable, but that synthetic generation which models content while ignoring layout variation will overestimate real-world accuracy. Any evaluation that relies on synthetic documents should treat generation fidelity—particularly layout fidelity—as a first-class experimental variable and report it, rather than treating synthetic documents as interchangeable with real ones.

5.4 Multi-Model Comparison

The three cloud models span 1pp on micro-average but diverge by up to 31pp per category: Claude leads GPT-4o-mini by 31pp on receipts and trails it by 21pp on medical records (Table 3). Local models trail the cloud tier but remain usable: Qwen3 8B reaches 83.3% micro—5.1pp below GPT-4o-mini—while running fully offline, and it leads GPT-4o-mini on receipts (60.5% vs. 57.8%). Mistral 7B and Llama 3 8B fall to 70.0% and 65.6%. For deployments where documents cannot leave the premises, the cost of staying local is therefore closer to 5pp than to the 20pp the weakest local models suggest.

The medical records regression (mini 86.2% vs. Claude 65.2%) is driven by **over-abstention**: frontier models return null on 92–105 fields where

mini correctly extracts (Tables 4–5), accounting for 83–97% of the gap. The affected fields have values stated verbatim in the source text (e.g., “Attending: Dr. A”), making GT effectively deterministic and the finding robust to annotation quality concerns.

5.5 Cross-Pipeline Comparison

We evaluate a full-document baseline (single-pass function calling, no routing) using the same model (GPT-4o-mini); Table 6 reports the comparison. The routed arm runs with no configurable techniques enabled (ablation step 0), routing on schema hints alone. The baseline is an independently built single-pass extractor, not a reconfiguration of our system, so the comparison reflects implementation as well as architecture.

The routed pipeline gains most on contracts (+26pp) and medical records (+23pp) and loses on claims (−5.5pp) and receipts (−6.1pp). Document length, as section count, does not explain this: across the ten categories the median is uncorrelated with the gap (Spearman $\rho = -0.05$), and medical records—the second-largest gain—are single-section documents. Length matters *within* categories, in opposite directions: across contract length quartiles the gap widens from +17.6pp to +35.2pp, while across insurance-claim quartiles it reverses from +6.1pp to −28.1pp. The contracts and legal filings results should be interpreted cautiously: the GT spot-check (Appendix F) found scope-boundary errors in 4/5 sampled contracts and enum-mapping errors in 5/5 legal filings, so absolute gaps may partially reflect GT noise. The *direction* is robust across categories. The most-failed contract fields are `governing_law` and `contract_type` (24% and 23% failure rates). Of the 92 contract field failures, 42 are wrong values, 38 are hallucinations on fields whose ground truth is absent, and 12 are missing values—so the dominant failure mode on long contracts is asserting a value where none exists, not declining to extract. The longest contracts (15–535 sections) score 76% vs. 86% for the shortest.

5.6 Technique Ablation

Context chunks (+3.2pp; Table 7) dominate; hints reduce accuracy by 1.5pp when added on top. The per-category impact varies dramatically (Table 11, Appendix C): receipts gain +21.8pp and invoices +7.9pp, where header sections supply the entity context that individual line items lack, whereas

Model	Type	Micro	Macro	Rept.	Med.	Inv.	Ctr.
4o-mini	Cloud	88.4	87.1	57.8 $\pm$ 8	86.2 $\pm$ 3	87.9 $\pm$ 2	81.6 $\pm$ 3
4o	Cloud	89.4	87.9	81.0 $\pm$ 6	70.3 $\pm$ 4	97.7 $\pm$ 1	76.8 $\pm$ 4
Claude Sonnet 4	Cloud	89.2	87.9	89.1 $\pm$ 5	65.2 $\pm$ 4	98.3 $\pm$ 1	74.0 $\pm$ 4
Qwen3 8B	Local	83.3	81.1	60.5 $\pm$ 8	69.5 $\pm$ 4	88.7 $\pm$ 2	69.2 $\pm$ 4
Mistral 7B	Local	70.0	69.7	44.2 $\pm$ 8	60.5 $\pm$ 4	75.4 $\pm$ 3	44.0 $\pm$ 4
Llama 3 8B	Local	65.6	65.7	37.4 $\pm$ 8	56.2 $\pm$ 4	70.0 $\pm$ 3	41.0 $\pm$ 4

Table 3: Accuracy by model (% , 95% Wilson CIs) across 6 models and 5 families. Cloud models span 1pp overall but diverge by up to 31pp per category. Local models run fully offline via Ollama.

Error type	Mini	4o	Claude
Missing (null, value exists)	23	130	146
Wrong value	44	34	51
Hallucinated	10	4	2
Format mismatch	6	10	10
Total failures	83	178	209

Table 4: Medical records errors (100 docs, 600 fields). “Wrong value” and “format mismatch” subdivide the non-null mismatches that §4’s four-way scoring counts as incorrect. Frontier models fail by *not extracting* (70–73%); mini by *wrong values* (53%). “Format mismatch” is a semantically-correct value in a differing format, scored as a failure.

Field	4o miss	Claude miss	Mini acc.
attending_physician	38	50	90%
medications_disch.	37	37	81%
procedures	8	6	70%
admission_date	4	8	92%
primary_diagnosis	3	2	89%
discharge_date	2	2	95%
Total	92	105	

Table 5: Over-abstention by field (frontier null, mini correct). MTSamples transcriptions use abbreviated notation that frontier models refuse to extract.

IRS forms gain nothing—their fields are already unambiguously positioned.

5.7 Routing Strategy Comparison

With context chunks enabled, we evaluate 8 routing conditions (Table 10, Appendix C; merge strategies defined in Appendix H).

No routing strategy improves over heuristic-only (88.9%); additive merge ties it exactly. Without context chunks enabled, additive merge gains +2.3pp, but that gain is a confound: the LLM map’s added sections happened to include document headers, providing the same benefit as context chunks. Once context chunks are controlled for, the LLM routing contribution is zero. Replacing heuristic with LLM-only routing loses 2.8pp.

Category	Routed	Full-doc	Δ	Sig.
IRS forms	100.0 $\pm$ 1	100.0 $\pm$ 1	0.0	—
Ins. policies	97.9 $\pm$ 1	88.2 $\pm$ 2	+9.7	✓
Ins. cert.	90.8 $\pm$ 3	79.0 $\pm$ 5	+11.8	✓
SEC filings	88.6 $\pm$ 1	86.6 $\pm$ 2	+2.0	—
Ins. claims	88.1 $\pm$ 2	93.6 $\pm$ 2	−5.5	✓ [†]
Invoices	87.9 $\pm$ 2	86.6 $\pm$ 2	+1.3	—
Legal filings	85.5 $\pm$ 3	73.3 $\pm$ 4	+12.2	✓
Med. records	86.2 $\pm$ 3	63.3 $\pm$ 4	+22.8	✓
Contracts	81.6 $\pm$ 3	55.6 $\pm$ 4	+26.0	✓
Receipts	57.8 $\pm$ 8	63.9 $\pm$ 8	−6.1	—
Micro / Macro	88.4 / 87.1	82.5 / 80.1	+5.9	✓

Table 6: Cross-pipeline comparison (same model: GPT-4o-mini) with 95% Wilson CIs. 6/10 categories significant, where significant means the two 95% Wilson intervals are disjoint—a stringent criterion (roughly $\alpha \approx 0.006$ per comparison) that is conservative under Holm correction across the 10 categories; intervals treat fields as independent and do not exploit the pairing between conditions. [†]The baseline’s win concentrates on multi-section claims.

Upgrading the mapper to GPT-4o (−0.2pp, 6 $\times$ slower) does not help. Schema hints contribute +1.8pp; LLM routing partially compensates when hints are absent (+1.0pp), suggesting value only when expert-authored hints are unavailable. Insurance certificates show the largest per-category sensitivity: heuristic 91.5% vs. additive 84.6%.

5.8 Per-Category Difficulty Analysis

Table 12 (Appendix C) gives per-category accuracy across routing conditions, revealing which categories are configuration-invariant and which are sensitive.

Under the heuristic-routing, context-chunks-enabled configuration of Table 12—not the no-techniques baseline of Table 2, against which per-category figures differ by up to 22pp—three difficulty tiers emerge. **Easy (>96%)**: IRS forms, policies, claims, invoices — rigid structure, routing-invariant (spread <1pp). **Medium (78–92%)**: Certificates, SEC, legal, receipts, medi-

Step	Configuration	Acc.	Δ_{prev}
0	Baseline	88.4%	—
1	+ Context chunks	91.5%	+3.2
2	+ Extraction hints	90.0%	−1.5
3	+ Gap-fill retry	90.2%	+0.3*
4	+ Wave extraction	90.3%	+0.1*

Table 7: Cumulative ablation (GPT-4o-mini). *Below ± 0.55 pp run variance. Deltas are conditional on prior techniques, not independent effects, and are computed on unrounded values.

cal, adversarial — moderate sensitivity (spreads 1–10pp). **Hard (<76%)**: Contracts — accuracy drops to 55.6% without routing and to 69.6% without hints.

6 Discussion

Pipeline architecture should be chosen per category. The natural explanation for the cross-pipeline gaps—that routing helps once a document is too long to attend over—is not what the data shows (§5.5). Median section count does not predict the gap across categories, and length modulates it *within* categories in opposite directions. What predicts the winner is the category, not the length, so architecture should be selected per document type rather than by a length threshold.

Context chunks are sequentially the dominant technique. The +3.2pp improvement costs no additional LLM calls, whereas gap-fill retry (+0.3pp for 33% more calls) and wave extraction (+0.1pp for 46% more calls) fall within run-to-run variance and are not cost-justified. With context chunks enabled, LLM-assisted routing adds zero value over schema-hint heuristics—LLM routing’s apparent gains are context chunks doing the work.

Null semantics change the metric. Without the four-way distinction, a system returning null everywhere scores well on categories with many conditionally-absent fields. Our adversarial category (93.3%) scores high precisely because most of its documents test graceful null-return—the correct behavior.

Model selection is domain-dependent. The multi-model comparison reveals that upgrading to a frontier model is not uniformly beneficial. Claude gains 31pp over GPT-4o-mini on receipts (89.1% vs. 57.8%) but *loses* 21pp on medical records (65.2% vs. 86.2%). The mechanism is

over-abstention, not capability: frontier models return null on fields where mini correctly extracts verbatim values. For heterogeneous document types, per-category model selection may therefore outperform a single frontier model deployed uniformly.

Real documents are essential, and the gap has an identifiable mechanism. The certificate gap (38–45pp depending on field-mix standardisation) shows that synthetic-only evaluation overestimates production accuracy, and the mechanism—template fidelity to layout rather than to content—predicts where such gaps should appear. We cannot test that broadly: only three categories hold both real and synthetic documents in usable numbers, and invoices (5 real) and IRS forms (0 real) admit no comparison at all. Evaluations should report real-vs-synthetic accuracy separately rather than assume the gap is uniform.

6.1 Practical Recommendations

For pipeline builders: start with context chunks (+3.2pp at zero additional LLM cost); choose architecture per category, not by document length—the routed pipeline gains 12–26pp on contracts, medical records and legal filings but costs 5–6pp on claims and receipts, and section count does not predict which ($\rho = -0.05$); use expert-authored schema hints for routing (+1.8pp) rather than LLM-assisted routing, which helps only when hints are unavailable; test model upgrades per category, since frontier models can regress 21pp on specific domains; and include real documents, since synthetic-only evaluation overestimated accuracy by 38–45pp here.

7 Conclusion

We present an empirical study of schema-driven document extraction across 1,114 documents, 12 categories, and 8 domains. Aggregate accuracy conceals divergence along every axis: three models score within 1pp overall but diverge by 31pp per category; two pipelines differ by 26pp on contracts; and real documents score 38–45pp lower than synthetic on insurance certificates. None of it is visible in the one number a practitioner would consult. The corpus and tooling are released as a pinned, citable snapshot (Appendix B).

Limitations

The 47 real certificate values are not a uniform sample of the schema: `policies` and `certificate_holder_name`, the two failing fields, are 64% of the real denominator but 33% of the synthetic one, and `producer_phone` has no real observations at all. Standardising to the synthetic field mix reduces the gap from 44.5pp to **38.2pp**; every one of the five fields annotated on both sides still shows a real-vs-synthetic deficit (20.0–58.3pp), so the effect is robust but its published magnitude is partly compositional.

The corpus is English-language business documents, weighted toward insurance (310/1,114). We test two pipeline architectures but not vision models or multi-agent systems. GT for non-index sources was seeded by GPT-4o-mini and reviewed by a single annotator; locked-protocol concordance (§3) supports the claim that GT is not calibrated to one implementation, and a blind second annotator agrees with GT on 81.9% of 371 fields across 60 stratified real documents. That is human verification of GT rather than two-annotator agreement established from scratch, and its weakest component— $\kappa=0.00$ on `governing_law`—marks enum schema definitions, not extraction, as the residual risk. The independent validation and GT seeding share one model (GPT-4o-mini); concordance from two additional non-seeding models (GPT-5.4-nano 82.1%, MiniMax M3 86.1%) substantially mitigates shared-model-bias concerns.

GPT-4o-mini both seeded non-index GT and serves as the baseline model, so its absolute accuracy may be inflated by correlated errors; comparisons among non-seeding models, and all within-model pipeline comparisons, are unaffected, and the source-grounding analysis (Appendix E) bounds fabricated values below 1% for medical records.

Length is operationalised as section count. Measured instead by document bytes the rank correlation with the cross-pipeline gap is +0.65 ($p=0.04$), driven by the smaller categories; but SEC filings, the longest category by bytes, still show only a +2.0pp gap, so no length measure predicts the winner in the sense the routing hypothesis requires. The cross-pipeline comparison and the routing null result both use a single extraction model, so whether either is model-invariant is untested; models with larger context windows may respond differently to context expansion.

The cumulative ablation (§5.6) reports conditional effects in one ordering only; §5.6’s $-1.5pp$ measures hints injected into the extraction prompt while §5.7’s $+1.8pp$ measures hints used to route; the two are distinct interventions on the same schema annotations. Four-way null semantics can inflate accuracy on categories with many absent fields; excluding the adversarial category changes the micro-average by 0.1pp (§4), indicating negligible impact. Run-to-run variance is $\pm 0.05pp$ corpus-level (§4).

Ethics Statement

The corpus contains documents from public, publicly accessible, or openly licensed sources: SEC EDGAR filings, which are mandatory public disclosures authored by the filing registrant and therefore not U.S. Government works; CourtListener and Caselaw Access Project opinions, which are uncopyrightable government edicts; SROIE under CC BY 4.0; MTSamples under its site terms; and forms published by federal, state and university bodies. Some public documents contain names and entities already present in the source records. No private or confidential documents are included. Synthetic documents are generated from templates, not from private data.

MTSamples transcriptions are *not* CC0. A CC0 label attached to a third-party mirror of the dataset does not originate from the rights holder and has no chain of title; MTSamples’ own terms permit sharing for educational purposes with credit. The site’s terms state that the sample reports may be printed, shared, or linked to for educational purposes, so we redistribute the medical subset under those terms with attribution to MTSamples as the origin, rather than as an open licence. Names and dates were altered by the source for confidentiality; we additionally scanned all 100 clinical notes for residual structured identifiers—social-security, telephone, email, and record numbers—and found none. Per-source licenses and redistribution modes for every category are listed in Appendix A.

Acknowledgements

Ground truth for non-index categories was seeded by GPT-4o-mini extraction and reviewed by a human annotator, as described in §3; the seeding model is also one of the systems evaluated, and the consequences are set out in Limitations. Gen-

erative AI assistance was used in preparing this manuscript for analysis of the scored output, for recomputation and verification of reported statistics, and for editing. All reported numbers were recomputed from the released scored output, and the authors take responsibility for the content.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*.
- Udi Barzelay, Ophir Azulai, Inbar Shapira, Idan Friedman, Foad Abo Dahood, Madison Lee, and Abraham Daniels. 2026. VAREX: A benchmark for multi-modal structured extraction from documents. *arXiv preprint arXiv:2603.15118*.
- E. F. Codd. 1990. *The Relational Model for Database Management: Version 2*. Addison-Wesley, Reading, MA.
- Docsumo. 2026. [Docsumo: Intelligent document processing](#). Product page. Advertises “99% field-level accuracy on 250+ document types”; no measurement protocol published. www.docsumo.com, accessed 2026-08-04.
- Doxis AI Solutions. 2026. [Doxis \(formerly Klippa\): Invoice OCR and data extraction](#). Product page. Advertises “invoice data extraction accuracy up to 99%”; no measurement protocol published. klippa.com/en/ocr/financial-documents/invoices, accessed 2026-08-04.
- Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. 2026. ExtractBench: A benchmark and evaluation methodology for complex structured extraction. *arXiv preprint arXiv:2602.12247*.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR 2019 competition on scanned receipt OCR and information extraction. In *International Conference on Document Analysis and Recognition*.
- Hyperscience. 2026. [Hyperscience: Hyperautomation platform](#). Product page. Advertises “accuracy rates of 99.5%”; no measurement protocol published. www.hyperscience.ai, accessed 2026-08-04.
- Guillaume Jaume, Hazım Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition*.
- 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*.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Winter Conference on Applications of Computer Vision*.
- Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A consolidated receipt dataset for post-OCR parsing. In *Document Intelligence Workshop at Neural Information Processing Systems (NeurIPS)*.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*.
- Anubhav Shrima, Aryan Jain, Soumyajit Chowdhury, and Promod Yenigalla. 2025. [PARSE: LLM driven schema optimization for reliable entity extraction](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2749–2763, Suzhou (China). Association for Computational Linguistics.
- Štěpán Šimsa, Milan Šulc, Michal Uříčář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. DocILE benchmark for document information localization and extraction. In *Document Analysis and Recognition – IC-DAR 2023*, volume 14188 of *Lecture Notes in Computer Science*, pages 147–166. Springer.
- Divyansh Singh, Manuel Nunez Martinez, Bonnie J. Dorr, and Sonja Schmer Galunder. 2025. SLIDE: Sliding localized information for document extraction. *arXiv preprint arXiv:2503.17952*.
- Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. 2025. [DocAgent: An agentic framework for multi-modal long-context document understanding](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17701–17716, Suzhou, China. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Conference on Computational Natural Language Learning*.
- Jordy Van Landeghem, Rubén Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty,

Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. 2023. Document understanding dataset and evaluation (DUDE). In *International Conference on Computer Vision*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. [LayoutLMv2: Multi-modal pre-training for visually-rich document understanding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. [LayoutLM: Pre-training of text and layout for document image understanding](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1192–1200. ACM.

A Corpus Sources, Licenses, and Redistribution

Table 8 lists every corpus source with its origin, license basis, and how we redistribute it. Every category ships in full except 35 insurance documents whose ACORD or ISO form layouts are copyrighted; those are distributed by reference (Appendix B). Where a source’s terms grant permission conditionally rather than under an open licence, we redistribute under those terms and attribute the origin.

B Per-Category Field Inventory

Corpus version. All results in this paper are computed against a single pinned corpus snapshot totalling 7,197 annotated field values across 1,114 documents. Later corpus releases change some annotations. We release this exact pinned snapshot alongside the paper so that every reported number is reproducible once the withheld documents below are re-fetched, and we version subsequent releases separately. It is deposited at doi.org/10.5281/zenodo.22135948, which is the citable form of the snapshot and the version these results were computed against. Thirty-five documents—18 ACORD certificate-of-insurance and 17 ISO policy-form samples, whose form layouts are copyrighted—are withheld from the deposit and distributed by reference: their ground-truth annotations and provenance manifests ship

with the archive, and each manifest records the publishing body’s own retrieval URL. These are all of the real insurance certificates and policies in the corpus, so any result involving them—including §5.3—requires re-fetching from the recorded URLs. The scorer implementing the methodology of §4—type-aware comparison and the four-outcome null rule—is released separately at github.com/fieldbench/fieldbench under an MIT licence; note that its default corpus is the later FieldBench release rather than the snapshot used here, so it must be pointed at the deposited snapshot to reproduce these numbers. Readers should cite that record rather than any later release: the successor corpus, published separately as FieldBench, has since diverged in both document membership and annotations, and its numbers will not match those reported here.

Table 9 reports, for each category, the number of documents, the number of distinct schema fields annotated, the total annotated field values, and the mean number of scored fields per document. Schemas define fields with explicit types—string, number, date, array, and enum—together with a natural-language description and optional extraction hints (category keywords, positional preferences). Field values whose ground truth is a correct absence are recorded as null and scored under four-way null semantics (§4); fields whose ground truth could not be established are omitted from the schema instance entirely and are therefore never scored.

C Per-Category Breakdowns

Table 11 reports the per-category effect of context chunks, the single highest-impact technique in the ablation (§5.6). Table 10 reports the full routing-strategy comparison summarized in §5.7. Table 12 reports per-category accuracy across the three routing conditions that differ materially, supporting the difficulty tiers in §5.8.

D Present-Fields-Only Accuracy

Table 13 reports accuracy restricted to fields with a non-null ground-truth value, alongside the headline figure that also credits correct absences (§4). Corpus-wide the two differ by only 0.3pp, so aggregate accuracy is not materially inflated by correct abstention. Within categories the gap is larger and runs in both directions: medical records, where 42% of scored fields are absent, falls 7.3pp

Category	Source	Retrieval URL	License basis	Redistribution
SEC filings, Contracts	SEC EDGAR	sec.gov/edgar	No explicit licence; mandatory public disclosure	Full
Legal filings	CourtListener	courtlistener.com	Public domain opinions	Full
Legal filings	Caselaw Access Project	case.law	CC0 / public domain	Full
Receipts	SROIE (ICDAR 2019)	rrc.cvc.uab.es/?ch=13	CC BY 4.0 (attribution)	Full (attribution)
Medical records	MTSamples	mtsamples.com	Site terms: educational use w/ credit	Full (attribution)
Insurance (real)	Govt. / state sites	per-doc URL in manifest	Published publicly; ACORD/ISO layouts copyrighted	By reference
Insurance, Invoices, IRS (syn.)	Generated by the authors	—	Authors’ own work	Full
Adversarial (25)	Generated by the authors	—	Authors’ own work	Full
Multi-format (3)	Govt. / state sites	per-doc URL in manifest	Publicly published documents	Full

Table 8: Corpus sources, retrieval URLs, license basis, and redistribution mode, covering all twelve categories. IRS forms are synthetic documents generated from public-domain federal form layouts, not retrieved filings; per-document retrieval URLs for the government-sourced insurance and multi-format documents are recorded in the corpus manifest.

Category	Docs	Fields	Values	/doc	Field types
SEC filings	288	7	2,016	7.0	string, date, enum
Invoices	155	7	1,085	7.0	string, date, number, array
Insurance claims	152	9	1,020	6.7	string, date, number
Contracts	100	5	500	5.0	enum, date, array
Medical records	100	6	600	6.0	date, string, array
Insurance policies	97	9	763	7.9	string, number, date
Ins. certificates	61	6	305	5.0	string, date, array
Legal filings	61	7	427	7.0	string, date, enum
Receipts	52	3	147	2.8	string, date, number
Adversarial	25	11	136	5.4	string, date, enum
IRS forms	20	17	180	9.0	number, string
Multi-format	3	6	18	6.0	string, date, number

Table 9: Per-category field inventory. “Fields” is the number of distinct annotated schema fields; “Values” is the total annotated field values scored in that category.

#	Strategy	Acc.	Δ
1	Heuristic-only	88.9%	—
2	LLM-map-only	86.1%	−2.8
3	Additive merge	88.9%	0.0
4	Intersection merge	87.2%	−1.7
5	Additive (GPT-4o map)	88.7%	−0.2
6	Embedding (TF-IDF)	86.9%	−2.0
7	Heuristic, no hints	87.1%	−1.8
8	Additive, no hints	88.1%	−0.8

Table 10: Routing comparison (GPT-4o-mini, context chunks enabled). No LLM routing strategy improves over heuristic-only.

Category	Step 0	+Context	Δ
Receipts	57.8%	79.6%	+21.8
Invoices	87.9%	95.9%	+7.9
Ins. claims	88.1%	95.0%	+6.9
Adversarial	93.3%	98.3%	+5.0
Legal filings	85.5%	86.9%	+1.4
SEC filings	88.6%	89.4%	+0.8
Contracts	81.6%	82.4%	+0.8
IRS forms	100.0%	100.0%	0.0

Table 11: Per-category context chunk impact. Categories where headers provide entity disambiguation benefit most; structured forms gain nothing.

under the present-only metric, whereas adversarial documents *rise*, because their failures are hallucinations on absent fields rather than misses on present ones. The effect is therefore not monotonic in absent-field share, which is why we report both.

E Source Grounding of Seeded Annotations

Because ground truth for non-index categories is seeded by model extraction, a natural concern is that annotations could encode the seed model’s over-extraction rather than the docu-

ment’s contents—which would reward systems that reproduce that behavior. We test this directly on medical records, the category where the seeded-GT concern is sharpest because it shows the largest model divergence (§5.4). Checking every ground-truth value against its source document—867 non-null values, counting array elements individually rather than the 600 flat field observations scored elsewhere—850 (98.0%) are present in the source text. Manual inspection of the 17 residual cases shows they are predominantly matching artifacts rather than fabrications: nominalized procedure names (“Foley catheter

Category	Heur.	Add.	Add. n-h	Spr.
IRS forms	100.0%	100.0%	100.0%	0.0
Ins. policies	98.8%	98.7%	98.2%	0.6
Ins. claims	97.5%	98.2%	98.5%	1.0
Invoices	96.3%	96.3%	96.2%	0.1
Ins. certificates	91.5%	84.6%	82.0%	9.5
Adversarial	87.5%	85.3%	86.0%	2.2
SEC filings	83.6%	85.0%	84.8%	1.4
Legal filings	82.9%	82.9%	82.2%	0.7
Receipts	80.3%	81.0%	75.5%	5.5
Med. records	78.7%	78.5%	77.0%	1.7
Contracts	75.4%	73.8%	69.6%	5.8
Overall	88.9%	88.9%	88.1%	0.8

Table 12: Per-category accuracy across routing conditions, all with context chunks enabled—so these figures differ from Table 2, which reports the no-techniques baseline. They also differ from Table 7’s step 1, which reports a separate run of the same nominal configuration; absolute levels are not comparable across the two, though within-table comparisons are. “Heur.” = heuristic-only, “Add.” = additive merge, “Add. n-h” = additive without hints. Spread = max – min across conditions. Insurance certificates and contracts show the largest configuration sensitivity.

placement” where the note reads “Foley catheter was placed”), one medication spelling variant, and date-format mismatches. True over-extraction is therefore below 1%—far too small to account for the 16–21pp model gap in this category. The fields on which frontier models abstain are genuinely present in the source, so that gap reflects extraction behavior (§5.4), not a ground-truth artifact.

F Blind Second-Annotator Agreement

A second annotator, blind to ground truth and working from the corpus annotation guidelines, re-annotated a stratified sample of 60 real documents (12 per category) drawn from the five real-document categories large enough to stratify. Synthetic documents are excluded: their ground truth is correct by construction and carries no agreement signal. Agreement is scored with the same type-aware comparison used for extraction (§4), so date and numeric normalisation apply equally to annotator and system.

Table 14 reports per-category agreement. It is ordered by schema ambiguity, not by document difficulty: SEC filings, whose fields are largely index-derived and unambiguous, reach 97.6%, while legal filings—where the annotator must map a free-text caption onto an enumerated `filing_type`—fall to 69.1%. Cohen’s κ on the four enumerated fields is `form_type`

Category	Headline	Present-only	Absent
Corpus	88.4%	88.1%	20%
Medical records	86.2%	78.9%	42%
Insurance claims	88.1%	86.9%	19%
SEC filings	88.6%	87.8%	37%
Insurance policies	97.9%	97.8%	4%
Ins. certificates	90.8%	90.9%	3%
Invoices	87.9%	87.9%	0%
Receipts	57.8%	57.8%	0%
IRS forms	100%	100%	0%
Legal filings	85.5%	88.0%	5%
Adversarial	93.3%	95.2%	65%
Contracts	81.6%	84.2%	32%

Table 13: Headline vs. present-fields-only accuracy (GPT-4o-mini baseline). “Absent” is the share of scored fields whose ground truth is null.

Category	Agreement	Partial
SEC filings	97.6%	97.6%
Receipts	85.9%	85.9%
Medical records	79.2%	80.9%
Contracts	76.7%	79.2%
Legal filings	69.1%	69.1%
Overall	81.9%	82.7%

Table 14: Blind second-annotator agreement with ground truth, 60 documents / 371 fields. “Partial” allows partial credit on array-valued fields. 67 disagreements total. A further 13 template fields where ground truth carries a value but the annotator left the field blank are excluded by the harness rather than scored as disagreements; counting them gives 79.2%.

1.00, `contract_type` 0.74, `filing_type` 0.63, `governing_law` 0.00, mean 0.59. The `governing_law` figure is a canonicalisation artifact rather than disagreement: in all twelve cases the annotator wrote “State of X” where ground truth reads “X”, identifying the same jurisdiction each time. Chance agreement is likewise zero because the two label sets are disjoint, which forces κ to 0. The defect is that the field lacks a canonical form, and the same surface mismatch inflates its reported extraction failures.

Disagreement adjudication. Manual adjudication of 100 disagreements sampled from the locked-protocol run (§3) found 51 extraction errors, 42 GT-side issues, 5 normalization mismatches, and 2 source ambiguities. Because the sample is conditional on disagreement, these percentages characterize *disagreement modes*, not corpus-wide GT quality. GT-side issues cluster into schema-scope patterns: enum ambiguity (legal filings), field-scope boundaries (contracts),

Category	Real	Synthetic	Gap
Ins. certificates	53.2±14	97.7±2	+44.5
Ins. policies	81.4±11	98.9±1	+17.5
Ins. claims	72.0±12	89.0±2	+17.0

Table 15: Real vs. synthetic accuracy (%; 95% Wilson CIs) for the three categories holding both. Real arms are 18, 17 and 22 documents (47, 43 and 50 annotated values); synthetic arms 43, 80 and 130 documents. The remaining nine categories are wholly real or wholly synthetic and admit no comparison.

period-field applicability (SEC filings), and clinical role distinctions (medical records).

Human GT spot-check. A stratified spot-check of 50 documents (5 per category, 10 categories) found 32 correct, 17 with at least one issue, and 1 ambiguous. GT errors concentrated in schema-ambiguous categories: legal filings (5/5 incorrect—enum mapping), contracts (4/5—scope boundaries), SEC filings (3/5—period-field scope). Unambiguous categories showed zero errors: invoices, IRS forms, insurance claims, and policies (all 5/5 correct).

G Real vs. Synthetic by Category

Three categories hold both real and synthetic documents in numbers that support a comparison. Table 15 reports all three. The direction is the same in every case and the magnitude varies with how far the generator reproduces the layout of the real form: insurance certificates, whose real documents are ACORD forms with unlabelled positional boxes and merged-cell grids, show the largest gap; claims and policies, whose real documents are closer in structure to their templates, show roughly 17pp each. Real arms are small—18, 22 and 17 documents—so the intervals are wide, and the certificate arm carries the field-mix caveat of §5.3. The consistency of the direction across three independent categories is the more robust observation.

H Experimental Setup

Techniques. *Context chunks* prepend the document’s opening header sections to every extraction group. *Extraction hints* inject the schema’s keyword and positional annotations into the extraction prompt. *Gap-fill retry* re-queries fields left null after the first pass. *Wave extraction* runs extraction in successive passes, each conditioned on

the previous pass’s output. In the routing comparison (§5.7), *additive merge* takes the union of the heuristic and LLM section selections and *intersection merge* their overlap.

All model comparisons use identical prompts, JSON structured output mode, temperature 0, and a 120-second timeout with retry on rate limits (exponential backoff: 2s, 4s, 8s). Model IDs: gpt-4o-mini-2024-07-18, gpt-4o-2024-08-06, claude-sonnet-4-20250514. The independent system uses gpt-4o-mini with function calling (single-pass, no routing); independent validation additionally used gpt-5.4-nano and minimax-m3-20260531 via OpenRouter. Cost and latency are not reported as they vary by provider pricing and infrastructure.