ONTOFORM: Ontology-Grounded Semantic Data Generation for Enterprise Document Understanding

Nilakshan Kunananthaseelan, Benjamin Kane, Siddharth Jain, Mahsa Ghorbanali, Fatemeh Shiri, Raj Tumuluri, Gholamreza Haffari
Openstream AI
{nilakshan,benjamin,siddharth,mahsa,fatemeh,raj,reza.haffari}@openstream.com

Abstract

Enterprise forms are layout-specific observations of real-world entities whose semantics span fields, recurring sections and related documents. Developing enterprise document-understanding systems requires large-scale supervision, but obtaining authentic forms is constrained by privacy concerns, annotation costs, and organization-specific formats. We present ONTOFORM, a semantic-first framework that uses a lightweight, task-specific ontology to generate enterprise form packets. The framework separates domain semantics from layout through ontology maps, hierarchical schemas, dependency-aware entity sampling, deterministic validation, and template bindings. It produces rendered forms with field boxes, ontology-keyed extraction targets, grounded question-answer pairs, and validation metadata. We instantiate the framework on three digitally rendered surrogate insurance form types, producing 706 forms. Finetuned vision-language models improve ontology-keyed extraction F_1 by 4.99 points and grounded-QA $ROUGE-L$ by 44.1 points over their zero-shot baselines. Ontology guidance further improves held-out-form extraction F_1 by 36.05 points overall and by 42.67 points on concepts absent from training, while increasing exact cross-form reasoning accuracy by 34.17 points. These results demonstrate the utility of the generated supervision signals for downstream tasks

1 Introduction

Structured enterprise forms are a central interface between business entities and operational decision-making systems. Insurance applications, loan packages, account openings, tax filings, customer onboarding, and government filings do not simply contain isolated fields. They encode observations about applicants, assets, exposures, histories, choices and attestations. Beyond simple extraction, systems must resolve cross-field relationships, ensure section-level value consistency, and align recurring entity details across multiple documents in a submission.

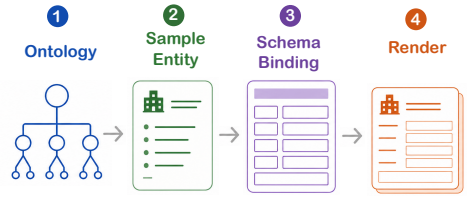


Figure 1: Overview of ONTOFORM: A latent enterprise entity is sampled under ontology constraints, bound to a template schema, and rendered as layout-specific documents.

Errors in enterprise document understanding can increase manual review time, delay intake and triage, and potentially introduce financial, compliance and operational risks. However, developing such systems at scale is constrained by privacy regulations (Regulation, 2016; Gramm, 1999), high annotation costs, and non-standardized vendor formats.

Existing document understanding benchmarks have made substantial progress on layout-aware extraction and visually rich document modeling (Jaume et al., 2019; Xu et al., 2022; Huang et al., 2022; Wang et al., 2023; Dua et al., 2025). However, public datasets typically treat documents as independent visual artifacts and provide limited support to explicit domain semantics, cross-field dependencies, or cross-document consistency. Synthetic document generators offer a practical path to scalable data creation (Kim et al., 2022; Lewis et al., 2006; Yim et al., 2021), but they often em-

phasize OCR robustness, visual diversity or type-valid field filling. Such documents may be visually plausible while containing inconsistent business attributes or failing to preserve semantic dependencies among related fields.

We present ONTOFORM, a semantic-first framework that uses a lightweight ontology map to generate digitally rendered, template-consistent enterprise form packets. ONTOFORM assumes that an enterprise form is a layout-specific observation of underlying business entities and their relationships. Therefore, generation is organized around three artifacts: a lightweight ontology map defining semantic concepts, value types, cardinalities, constraints and dependencies; a hierarchical schema binding ontology paths to template-specific fields and locations; and a sampler that generates coherent semantic records before rendering.

This paper makes three contributions: **(1)** a semantic-first generation framework that separates a lightweight, task-specific ontology and coherent entity sampler from document layout; **(2)** a controlled surrogate insurance instantiation spanning three related form types, with traceable field annotations, ontology-keyed targets, grounded QA, and validation metadata; and **(3)** controlled evidence from stronger architecture baselines, matched ontology ablations, multi-form reasoning, and held-out form-type transfer.

2 Related Works

Document Understanding. Form and visually rich document benchmarks have advanced semantic entity extraction, key-value linking, layout-aware modeling, and structured prediction. FUNSD (Jaume et al., 2019) introduced entity labeling and link prediction for scanned forms, and XFUND (Xu et al., 2022) extended form understanding to multilingual settings. Recent resources such as Form-NLU (Ding et al., 2023), VRDU (Wang et al., 2023), and DocILE (Šimsa et al., 2023) provide complex layouts, richer schemas, and business-document extraction tasks. These benchmarks primarily supervise document-local extraction from fixed datasets. In contrast, ONTOFORM treats each form as a projection of an ontology-grounded entity record, enabling semantically connected documents, ontology-keyed supervision, and consistency checks across related fields and forms.

Synthetic Document Generation. Synthetic document generation reduces dependence on private or expensive real data and has been useful for OCR, recognition, and document pretraining. SynthTIGER (Yim et al., 2021) generates synthetic text images for recognition, while Donut (Kim et al., 2022) uses synthetic document images for OCR-free document understanding. FlexDoc (Dua et al., 2025) is using schema-based synthetic documents to reduce privacy and annotation barriers. However, existing synthetic pipelines prioritise visual diversity, recognition robustness, or schema-valid document generation rather than semantic consistency across related records. ONTOFORM instead samples a coherent latent entity and projects it into multiple forms while preserving canonical concept identities, cross-field dependencies, and cross-document invariants.

Schema-Guided Tasks. Schema-guided extraction methods use structured target schemas or ontologies to constrain model outputs and improve extraction consistency (Cho et al., 2026; Cotti et al., 2026). These approaches generally operate over existing document collections, using schema information to guide prediction rather than to construct the underlying training data. ONTOFORM uses semantic structure earlier in the pipeline: the ontology defines the entities, attributes, dependencies, and constraints from which coherent records and their corresponding documents are generated. This enables schema-grounded supervision to be produced together with the documents themselves.

3 ONTOFORM

ONTOFORM treats the business entity, rather than the document, as the primary object of generation. Each enterprise form is then a layout-specific projection of an ontology-grounded entity record. Unlike synthetic pipelines that populate fields independently, this representation allows shared identities, attributes, exposures, schedules, policy details, and coverage selections to remain consistent across related forms.

The input is a pair $(\mathcal{O}, \mathcal{T})$, where $\mathcal{O}$ is an ontology map for a document family and $\mathcal{T}$ is a form template. First, ONTOFORM compiles the semantic concepts, value types, cardinalities, constraints, and template bindings into a hierarchical schema. It then samples a coherent latent entity record according to the ontology dependencies and validates the resulting record. Finally, the renderer projects

the validated record into $\mathcal{T}$ and exports the corresponding ground truths, including bounding boxes, ontology-keyed extraction targets and grounded QA pairs. Figure 2 summarizes the framework.

3.1 Ontology

We use a lightweight task-specific semantic representation rather than a full OWL implementation. It defines the layout-independent semantics of a document family. We formally define this ontology representation as a typed structure:

$$\mathcal{O} = (\mathcal{C}, \mathcal{P}, \mathcal{R}, \mathcal{K}), \tag{1}$$

where $\mathcal{C}$ is the set of concept classes, $\mathcal{P}$ is the set of associated properties with those concepts, $\mathcal{R}$ specifies the semantic relations among concepts, and $\mathcal{K}$ is the set of constraints, dependencies or sampling rules. In the insurance domain, concepts include applicants, locations, operations, fleet profiles, exposures, coverages, and loss histories. Properties include values such as legal name, industry category, payroll, coverage period, limits and deductibles.

$\mathcal{O}$ defines the semantic concepts represented by document fields and their relationships independently of visual layout. This distinction is critical for enterprise forms where the shared semantic concept may appear under different labels, positions, or grouping conventions across carrier-specific templates. For example, the legal name of the applicant or the mailing address may be filled differently across submission forms but must remain linked to stable entity values.

We group dependency-sensitive attributes, such as business size, industry-specific metrics, or exposure indicators, into subclasses linked to the primary entity. Enforcing this causal structure allows the sampler to generate contextually correlated attributes independently while keeping static identifiers (e.g., legal names, tax IDs, classification codes) factorized, ensuring joint statistical validity across the synthetic data.

For implementation, we realize each document-family ontology map as a JSON/YAML-compatible tree. The nodes represent semantic concepts and repeated concepts; leaf nodes declare value type, required status, cardinality, dependencies, choices and validation rules. The present system does not perform description-logic reasoning, OWL classification, or ontology alignment. This representation is more structured than an ordinary output schema because it is shared across templates and drives sampling, validation, rendering, and supervision, but it is less expressive than a general-purpose formal ontology. Nevertheless, it provides a practical semantic representation for reproducible data generation.

3.2 Hierarchical Schema and Layout Binding

The hierarchical schema is the bridge between form semantics and layout-specific properties. Given an ontology map $\mathcal{O}$ and template $\mathcal{T}$, the compiler produces a schema $\mathcal{S}$ whose internal structure reflects ontology concepts and binds ontology paths to template fields. For each leaf binding i, we represent the schema entry as

$$\mathcal{S}_i = (\pi_i, \tau_i, r_i, d_i, f_i, p_i, b_i), \tag{2}$$

where π_i is the ontology path, τ_i is the value type, r_i indicates whether the field is required, d_i stores dependency or constraint references, f_i is the template field identifier, p_i is the page index, and b_i is the normalized bounding box coordinates.

The hierarchical schema decouples semantic generation from layout-specific properties. The entity sampler operates over ontology paths and constraints, while the document renderer uses the schema to project the sampled record into the correct visual locations. This separation improves maintainability; adding a new carrier template requires updating ontology-to-field bindings, while the ontology, validation rules, and entity sampler can be reused.

3.3 Entity Sampling and Validation

ONTOFORM samples a latent entity record before rendering any document. Let z denote the entity record and s denote an optional scenario, such as business type, size, or line of business configuration. ONTOFORM uses the dependency declarations in $\mathcal{O}$ to construct a generation order over fields, which specifies the topological order in which the concepts and their properties should be sampled.

We denote the dependency structure as $G = (V, E)$, where each node V denotes an ontology field and each edge E denotes the declared sampling dependency.

$$p(z \mid s) = \prod_{v \in \mathrm{order}(G)} p(z_v \mid z_{\mathrm{pa}(v)}, s), \tag{3}$$

where $\mathrm{pa}(v)$ denotes the declared parent variables of v. Each conditional term $p(z_v \mid z_{\mathrm{pa}(v)}, s)$

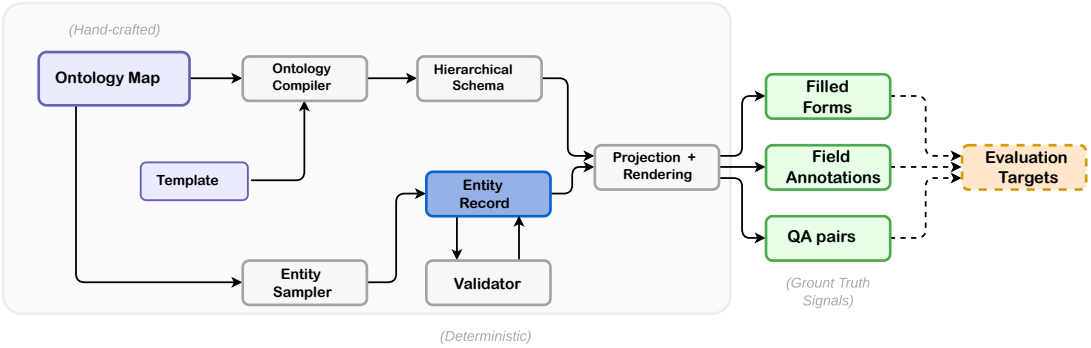


Figure 2: The ONTOFORM pipeline. Ontology maps and surrogate templates are compiled into hierarchical schemas. The dependency-aware samplers generate coherent business records, which are validated, rendered into forms, and exported with ontology-keyed field annotations.

is implemented by a field-specific sampler declared in the ontology map, using rule-based logic or simple parametric distributions conditioned on previously sampled values.

After sampling, a deterministic validator $V_{\mathcal{K}}(z)$ checks the record against constraints in $\mathcal{K}$. These include type checks, required-field checks, cardinality constraints, choice-set membership, temporal constraints, and cross-field or cross-document invariants. A candidate record is accepted only if $V_{\mathcal{K}}(z) = 1$. If validation fails, ONTOFORM resamples the violating subtree when possible rather than discarding the entire record.

For multi-form submissions, the same latent entity record is shared across document-specific projections. This enforces agreement for shared concepts such as applicant identity, mailing address, policy dates, locations, and scheduled assets, while allowing form-specific samplers to add line of business details.

3.4 Data Rendering and VLM Supervision Signals

The renderer projects the validated entity record into the template according to the hierarchical schema. For each schema leaf, the renderer retrieves the value associated with the ontology path, normalizes it for the target field type, and writes it into the bound template field. The same schema also determines which fields are absent, optional, repeated, or conditionally rendered.

Because ONTOFORM controls both the semantic record and its document projection, it exports exact supervision without manual per-example annotation. For each rendered form, the framework produces filled PDFs, page images, per-field bounding boxes, ontology-keyed extraction targets, grounded QA pairs, and validation metadata. These artifacts support local field extraction, ontology-path extraction, grounded QA, abstention evaluation, cross-field consistency checking, and cross-document agreement checking. They make ONTOFORM a scalable, semantic-first framework for grounded enterprise form understanding.

4 Insurance Instantiation and Evaluation Corpus

We instantiate ONTOFORM in commercial insurance because it exhibits the complex enterprise-document properties, including privacy-sensitive submissions, heterogeneous carrier-specific layouts, repeated entities, cross-document variation, and shared entity properties, that motivate our framework. We use these surrogate templates to preserve the structural and semantic challenges of insurance application workflows.

4.1 Surrogate Templates

We create three fillable surrogate templates covering general application, a general liability application, and a business auto application. The commercial application captures applicant identity, business operations, exposure summaries, locations, coverage interests, additional-document status, and attestation. The general liability form adds premises and operations exposure, liability limits, products and completed operations, subcontractor controls, prior coverage, liability losses, and attestation. The business auto form adds fleet structure, garaging locations, vehicle and driver schedules, coverage limits, and loss history. Together, these templates require models to handle typed fields, checkboxes, repeated groups, tables, section-specific meanings, and cross-form consistency.

Form	Inst.	Avg. fields	Ann.	QA
Commercial app	253	66.82	16,906	1,389
General liability	253	89.46	22,634	2,024
Business auto	200	118.89	23,778	3,982
Total	706	–	63,318	7,395

Table 1: Evaluation corpus details. Inst. counts rendered form instances; Ann. counts per-field annotations; QA counts grounded question-answer pairs.

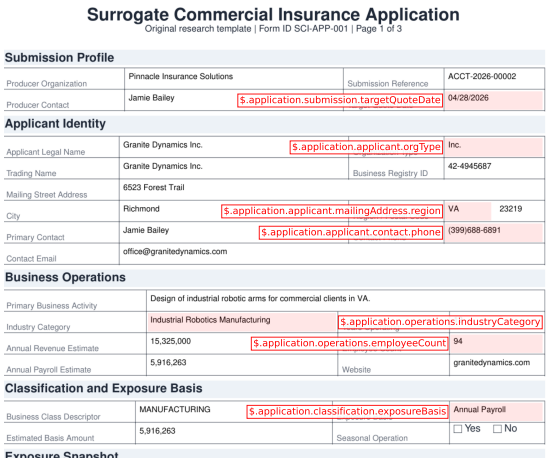


Figure 3: Example form instance with annotated ontology-path callouts (red). Extracted fields map to ground-truth values and pixel-level bounding boxes recorded during data generation.

4.2 Evaluation corpus

For downstream evaluation, we instantiate a surrogate insurance corpus from shared latent `business_entity` records and one or more lines of business. The shared entity captures applicant-level attributes such as identity, address, operations, exposure, and policy context, while the sampler adds relevant line of business details. For each rendered form, ONTOFORM exports the ontology-grounded payload, compiled hierarchical schema, field annotations, grounded QA pairs, and validation metadata. All artifacts are generated from the same source record, making each rendered value traceable to an ontology path, sampling rule, and template location. The details of the evaluation corpus are provided in Table 1. Figure 3 illustrates a page from a rendered submission annotated with ontology-path callouts. Every visible field carries a stable ontology path, ground-truth value, and pixel bounding box recorded at generation time. We also provided intrinsic validation of the corpus in Appendix A.

5 Empirical Validation

We evaluate ONTOFORM as a controlled feasibility study. All primary training and test documents are digitally rendered surrogate forms, allowing semantic records, rendered values, schema bindings, and supervision targets to be observed exactly. Entity-level splits prevent the same sampled business profile from appearing in training and test sets. The held-out-form protocol additionally removes one form type from training, but it remains an evaluation within the surrogate generation pipeline. Consequently, we focus on three questions: **(Q1)** Does ONTOFORM generate structurally valid and semantically coherent form instances? **(Q2)** Do generated annotations improve downstream tasks? **(Q3)** Does ontology grounding provide measurable benefits beyond independent synthetic field filling?.

The controlled experiments validate transfer of synthetic supervision rather than performance on authentic real-enterprise forms. This allows us to measure properties that are exactly observable from the generator: ontology coverage, template binding correctness, semantic consistency, rendering validity, grounded supervision quality, and downstream extraction usability.

For the downstream tasks, we finetuned Qwen3.5 VLMs (Bai et al., 2025) and compared with other baselines Additional training details are provided in Appendix B.1.

5.1 Grounded QA and Ontology-Keyed Extraction

We evaluate ONTOFORM supervision data in two downstream document-understanding tasks: grounded question answering and ontology-keyed JSON extraction. These tasks test whether the corpus supports both structured field extraction and localized semantic supervision, rather than only visual form rendering.

Grounded QA Each example consists of a rendered form context, a question generated from the ontology-grounded payload, and a normalized reference answer. Table 2 shows that finetuning on ONTOFORM data (LoRA FT) substantially improves grounded QA performance over the zero-shot baseline (ZS). The improvement is especially important because the questions are keyed to ontology-derived supervision rather than only natural-language surface labels. Additional details on training and evaluation are available in Appendix B.2.

Model	BLEU	METEOR	ChrF	R-1	R-2	R-L
Zero Shot	1.6	12.6	18.1	53.1	8.6	53.1
FT–w/o $\mathcal{O}$	62.6	54.4	81.8	97.2	15.9	97.2

Table 2: Overall grounded QA performance for Qwen3.5-2B. Scores are reported as percentages after answer normalization. R-1, R-2, and R-L denote ROUGE-1, ROUGE-2, and ROUGE-L.

Training	Inference	Exact	Prec.	Recall	Acc.+TN	F_1
Donut	×	75.65	98.77	75.65	–	85.68
Zero-shot	×	2.87	92.74	4.84	2.89	9.21
Zero-shot	✓	**54.11**	83.71	**91.33**	83.64	87.36
FT-w/o $\mathcal{O}$	×	47.53	**99.03**	80.22	75.17	88.63
FT-w/ $\mathcal{O}$	✓	51.31	98.90	86.61	**90.40**	**92.35**

Table 3: Ontology-guided extraction performance on 117 samples containing 5,784 fields (%). "Acc.+TN" includes correctly omitted null fields. Donut is evaluated only on populated annotated fields.

5.2 Ontology Guidance

We first separate the benefit of finetuning from the benefit of ontology guidance. In the ontology-guided setting, we use the ontology paths, semantic types, allowed values and schema constraints in addition to the images, base model, training budget, and field values. We additionally fine-tune Donut (Kim et al., 2022) as an architecture baseline using image-annotation pairs.

Donut obtains higher exact match but lower recall and F_1 than the finetuned VLM, providing complementary evidence that the generated image-annotation pairs are useful across model architectures. Compared with matched fine-tuning without ontology supervision, ontology-guided fine-tuning improves exact-match accuracy by 3.78 points, recall by 6.39 points, accuracy including true negatives by 15.23 points, and F_1 by 3.72 points. Ontology information also produces a large zero-shot gain, but the best balanced performance is obtained when ontology guidance is used consistently during finetuning and inference. Detailed generator-corruption ablations are provided in the Appendix B.3.

5.3 Controlled Generalization to Unseen Forms and LOBs

We evaluate form-type generalization under two complementary settings. First, we hold out one form type from the original surrogate pipeline, providing a controlled test of transfer to a new layout and field organization. Second, we evaluate on two unseen lines of business whose form structures and LOB-specific concepts are absent from training.

	Exact match		F_1		
Scope	w/o $\mathcal{O}$	w/ $\mathcal{O}$	w/o $\mathcal{O}$	w/ $\mathcal{O}$	ΔF_1
All fields	26.73±3.63	**58.13±2.50**	60.90±5.75	**96.95±2.52**	**+36.05**
Shared concepts	88.46±12.09	**99.15±0.60**	93.45±6.52	**99.15±0.60**	**+5.70**
Held-out concepts	21.23±2.88	**54.48±2.78**	53.92±5.37	**96.59±3.03**	**+42.67**

Table 4: Extraction on the held-out form type.

Throughout this section, w/o $\mathcal{O}$ denotes no-ontology fine-tuning with the no-ontology prompt, whereas w/ $\mathcal{O}$ denotes ontology fine-tuning with the ontology-conditioned prompt.

Held-out form within the original family. We train on the Commercial Application and General Liability forms and evaluate on the held-out Business Auto form. Shared concepts occur in at least one training form, whereas held-out concepts occur only in the Business Auto form.

As shown in Table 4, ontology guidance improves extraction F_1 by 36.05 percentage points across all fields and by 42.67 points on concepts absent from the training forms.

Transfer to unseen surrogate LOBs. The unseen-LOB evaluation uses 200 matched business-entity profiles shared across the two target LOBs. Each profile is rendered once per LOB as a four-page form, yielding 200 forms and 800 pages for Workers' Compensation and the Commercial Property.

Table 5 shows that ontology guidance improves canonical path-value F_1 by 87.3 points on Workers' Compensation and by 75.0 points on Commercial Property. The key-agnostic improvements are 39.6 and 46.9 points, indicating that ontology guidance primarily strengthens semantic path alignment while also improving value recovery. Additional details in Appendix B.3.

5.4 Multi-Form Reasoning

Because one sampled entity is rendered consistently across multiple related forms, we test whether a model can combine evidence across an application packet to answer underwriting questions. Each condition is evaluated over three seeds with 40 questions per seed. Training entities are disjoint from test entities.

Across three seeds, ontology-guided finetuning without inference-time ontology instructions achieves the highest macro-F_1 (90.80) and exact-reasoning accuracy (52.50), exceeding no-ontology

Unseen LOB	w/o $\mathcal{O}$	w/ $\mathcal{O}$	ΔF_1
Canonical path–value F_1			
Workers' Compensation	4.1 ± 1.2	$\mathbf{91.4 \pm 1.5}$	$\mathbf{+87.3}$
Commercial Property	5.2 ± 4.4	$\mathbf{80.2 \pm 14.8}$	$\mathbf{+75.0}$
Macro-average	—	—	$\mathbf{+81.2}$
Key-agnostic value F_1			
Workers' Compensation	52.6 ± 0.7	$\mathbf{92.2 \pm 0.7}$	$\mathbf{+39.6}$
Commercial Property	47.8 ± 5.2	$\mathbf{94.7 \pm 1.3}$	$\mathbf{+46.9}$
Macro-average	—	—	$\mathbf{+43.3}$

Table 5: Extraction on unseen surrogate form types. The canonical metric requires prediction of the correct ontology path and value, whereas the key-agnostic metric evaluates value recovery independently of the predicted path.

FT	$\mathcal{O}$ Inference	Macro F_1	Answer Acc.	Fact F_1	Exact Reason.
×	×	82.83	83.33	79.78	18.33
×	✓	72.89	75.00	78.99	15.00
✓	×	**90.80**	**90.83**	88.93	**52.50**
✓	✓	83.43	84.17	**89.98**	50.00

Table 6: Cross-form reasoning results. Exact reasoning measures exact agreement with the required structured reasoning facts.

finetuning by 7.97 and 34.17 points, respectively. For the ontology-finetuned model, adding ontology instructions at inference increases reasoning-fact F_1 from 88.93 to 89.98 but lowers macro-F_1 from 90.80 to 83.43.

6 Discussion

The empirical evidence suggests that ONTOFORM is useful, because it renders complex documents at scale, while preserving the alignment among ontology paths, entity values, and layouts. This distinction matters in enterprise settings, where field values are rarely independent, and the same business concept may recur across fields, forms, and workflows in different contexts. By making these links explicit, ONTOFORM provides ground ,truths for models to learn canonical semantic targets rather than only surface labels or coordinates.

6.1 Authoring and Maintenance Cost

The principal adoption cost is the one-time construction and validation of the ontology map, dependency rules, validators, and template-field bindings. Table 7 separates AI/API cost from estimated human labour; the estimates assume labour rates of US$75-150 per hour and should not be interpreted as deployment measurements. Adding a new carrier template still requires a new template-field binding, while changes to shared concepts, dependency rules, and validators can be reused across templates.

Authoring method	Human time	AI/API	Est. labour
AI draft + concept check	1–3 h	$1–20	$75–450
AI draft + domain validation	5–12 h	$1–20	$375–1,800
Fully manual	12–30 h	–	$900–4,500

Table 7: Estimated one-time authoring cost for the present three-form instantiation.

7 Limitations

This study evaluates ONTOFORM using digitally rendered, template-consistent surrogate insurance forms. The held-out-form and unseen-LOB experiments introduce layouts, schemas, and concepts absent from training, but remain within a controlled synthetic pipeline. They do not demonstrate transfer to authentic carrier documents or to scans containing OCR errors, handwriting, stamps, skew, low resolution, missing fields, or evolving layouts. Evaluation on de-identified or manually redacted enterprise forms remains important future work.

ONTOFORM uses a lightweight, task-specific ontology rather than a complete OWL implementation. It represents concept hierarchies, typed properties, constraints, dependencies, and template bindings, but does not perform description-logic reasoning or OWL classification. Developing and maintaining these artifacts requires domain expertise and may introduce template-specific biases. Finally, strong performance on synthetic data does not establish that an automated system is safe for consequential insurance decisions. Production use would require human review, monitoring for distribution shift, and evaluation of fairness and unequal error rates across relevant groups.

8 Conclusion

Enterprise form understanding should not be reduced to extraction from isolated pages. ONTOFORM shows that enterprise forms can instead be generated and supervised as layout-specific observations of ontology-grounded entities. This semantic-first view enables traceable synthetic supervision without production documents and offers a practical path toward scalable, privacy-preserving document-understanding systems for insurance and other regulated enterprise domains.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Nicole Cho, Kirsty Fielding, William Watson, Sumitra Ganesh, and Manuela Veloso. 2026. Taser: Table agents for schema-guided extraction and recommendation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 226–252.

Luca Cotti, Idilio Drago, Anisa Rula, Devis Bianchini, and Federico Cerutti. 2026. Ontologx: Ontology-guided knowledge graph extraction from cybersecurity logs with large language models. *Advanced Intelligent Systems*, page e202501381.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Form-nlu: Dataset for the form natural language understanding. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2807–2816.

Karan Dua, Hitesh Laxmichand Patel, Puneet Mittal, Ranjeet Gupta, Amit Agarwal, Praneet Pabolu, Srikant Panda, Hansa Meghwani, Graham Horwood, and Fahad Shah. 2025. FlexDoc: Parameterized sampling for diverse multilingual synthetic documents for training document understanding models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1500–1521, Suzhou (China). Association for Computational Linguistics.

Phil Gramm. 1999. Gramm-leach-bliley act. *Vol. Public Law*, pages 106–102.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 665–666, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–3.

Štěpán Šimsa, Milan Šulc, Matyáš Skalickỳ, Yash Patel, and Ahmed Hamdi. 2023. Docile 2023 teaser: document information localization and extraction. In *European Conference on Information Retrieval*, pages 600–608. Springer.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. Xfund: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the association for computational linguistics: ACL 2022*, pages 3214–3224.

Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. 2021. Synthtiger: Synthetic text image generator towards better text recognition models. In *International Conference on Document Analysis and Recognition*, pages 109–124. Springer.

A VLM-Based Quality Assessment

We complement the deterministic corpus checks with a model-based assessment of the visible quality of generated multi-form bundles. The deterministic checks verify ontology-binding coverage, field geometry, payload types, temporal constraints, and cross-form identity constraints given in 8.

The VLM assessment instead examines whether the rendered documents appear semantically coherent, mutually consistent, answerable, and useful as downstream supervision. We use a frozen Qwen3.5-4B model as the assessor. It receives matched multi-form bundles without being told which generation condition produced them. For each bundle, the model returns structured JSON containing a score from 1 (poor) to 5 (best) and supporting evidence for four dimensions:

- **Within-form coherence**: whether fields within each form describe a plausible and internally consistent entity;
- **Cross-form coherence**: whether shared facts agree across forms belonging to the same application;
- **Evidence answerability**: whether the rendered forms contain sufficient evidence to answer document-grounded questions; and
- **Downstream utility**: whether the bundle appears suitable as supervision for extraction and multi-form reasoning.

The composite score is the unweighted mean of these four dimensions. The assessor also returns a binary decision indicating whether a bundle is usable for training. Table 9 reports the average absolute scores over 200 matched entity bundles.

Layer	Measure	Score
E1	Schema/ontology binding coverage	506/506 leaves
E2	Layout/bbox geometry pass rate	506/506 widgets
E3	Type + cross-form invariants	76,134/76,134
E4	Generation throughput	0.73 subs/s
E5	VLM judge agreement	57.7% (F1 0.78)
E6	VLM judge robustness retention	95.6%

Table 8: Corpus validation. E1-E4 are deterministic checks over templates, ontology bindings, geometry, constraints, and throughput. E5 and E6 are secondary VLM-as-judge diagnostics for visual legibility and robustness.

As shown in Table 9, the ontology-grounded bundles receive consistently high scores across all four dimensions. Random filling and field-value mismatch produce individually implausible forms and therefore receive scores near 1. Cross-form contradictions are less severe within each individual form, obtaining a within-form score of 3.10, but their cross-form score falls to 1.16 because shared entity facts conflict across documents. This distinction indicates that the assessor responds specifically to cross-document inconsistency rather than only to local field plausibility.

Because absolute model-generated scores can be sensitive to calibration, we also perform paired comparisons. Each ontology-grounded bundle is compared with a baseline bundle generated from the same underlying entity. For entity i and baseline b, the paired composite improvement is

$$\Delta_{i,b} = S_{\text{ontology},i} - S_{b,i},$$

where S denotes the composite quality score. We report the mean paired difference, a 95% confidence interval obtained by entity-level bootstrap, and the number of ontology-grounded wins, ties, and losses. Results are presented in Table 10.

Table 10 confirms that the absolute-score pattern is also present under entity-matched comparison. Ontology-grounded generation outperforms random filling in all 199 valid pairs and field-value mismatch in all 196 valid pairs. Against cross-form contradictions, it wins in 195 of 200 pairs and ties in the remaining five, with a mean composite improvement of 2.598 points. The largest qualitative distinction is cross-form coherence: contradictory bundles can remain locally plausible while failing to represent one consistent business entity.

These results are supplementary model-based evidence. The baselines are deliberately corrupted controls, and the assessor is itself a VLM rather than a human evaluator. We therefore use the paired comparisons as a diagnostic of semantic consistency and retain deterministic validation and downstream task metrics as the primary evidence.

B Training and Evaluation Details

B.1 Training Configuration

The configuration used for extraction training is provided in the Table 11.

(a) Commercial Application (b) Business Auto (c) General Liability

Figure 4: Examples of the three digitally rendered surrogate insurance form types used in the controlled evaluation.

Generation condition	Within-form	Cross-form	Answerability	Utility	Composite	Training-usable
OntoForm	5.00	5.00	5.00	5.00	5.00	100%
Random field filling	1.01	1.00	1.01	1.01	1.00	0%
Field-value mismatch	1.00	1.00	1.00	1.00	1.00	0%
Cross-form contradictions	3.10	1.16	3.13	2.23	2.40	11%

Table 9: Absolute frozen-VLM quality assessment of 200 matched multi-form bundles. Each dimension is scored from 1 (poor) to 5 (best). Composite is the unweighted mean of the four dimension scores. Training-usable is the percentage of bundles that the assessor classifies as suitable for downstream supervision.

B.2 Grounded QA

Data Preparation We evaluate three ontology-grounded question types. **Extractive** questions ask for values explicitly rendered in a visible field, such as an applicant name, phone number, employee count, or policy date. **Boolean** questions ask about binary fields, selected options, or yes/no conditions. **Grounding** questions require the model to associate an answer with field-level evidence, such as the rendered field.

Metrics. For each prediction $\hat{y}$ and reference answer y, we apply the same type-aware normalization before scoring. Normalization includes lowercasing where appropriate, whitespace normalization, date normalization, currency normalization, and canonicalization of boolean values. We report BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ChrF (Popović, 2015), and ROUGE-1/2/L (Lin, 2004).

Let Q be the set of evaluation questions and let $\tilde{y}_i = \mathrm{norm}(y_i)$ and $\tilde{\hat{y}}_i = \mathrm{norm}(\hat{y}_i)$ denote the normalized reference and prediction for question i. For any metric m, the reported macro score is

$$\mathrm{Score}_m = \frac{1}{|Q|} \sum_{i \in Q} m(\tilde{\hat{y}}_i, \tilde{y}_i). \tag{4}$$

For question-type analysis, let $Q_t \subset Q$ denote the subset of questions of type t. We report

$$\mathrm{Score}_{m,t} = \frac{1}{|Q_t|} \sum_{i \in Q_t} m(\tilde{\hat{y}}_i, \tilde{y}_i). \tag{5}$$

The individual lexical metrics are computed as follows. ***BLEU*** is the clipped n-gram precision with brevity penalty:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right), \tag{6}$$

where p_n is clipped n-gram precision, w_n are uniform weights, and

$$\mathrm{BP} = \begin{cases} 1, & c > r, \\ \exp(1 - r/c), & c \le r, \end{cases} \tag{7}$$

with prediction length c and reference length r.

METEOR is computed as a harmonic mean of unigram precision and recall with a fragmentation

Paired baseline	Valid pairs	Composite Δ [95% CI]	Wins/Ties/Losses
Random field filling	199	3.996 [3.989, 4.000]	199/0/0
Field-value mismatch	196	4.000 [4.000, 4.000]	196/0/0
Cross-form contradictions	200	2.598 [2.516, 2.675]	195/5/0

Table 10: Entity-paired improvement of ontology-grounded bundles over each baseline. Composite Δ is the ontology-grounded composite score minus the baseline score for the same entity; positive values favor ontology-grounded generation. Pair counts differ when a condition lacks a valid structured VLM judgment.

Component	Setting
Base model	`unsloth/Qwen3.5-2B` (Daniel Han and team, 2023)
Train examples	1,698 pages
Validation examples	204 pages
Test examples	216 pages
Fine-tuning method	PEFT (Mangrulkar et al., 2022) LoRA (Hu et al., 2022) adapters
LoRA rank	$r = 16$
LoRA alpha	$\alpha = 16$
LoRA dropout	0.0
Adapted modules	Attention Q/K/V/O projections and MLP up/down/gate projections
PEFT version	0.19.1
Hardware	Single NVIDIA A5500 24GB

Table 11: Training setup for the ontology-guided VLM extraction model. Examples are page-level samples, not submission-level packages.

penalty:

$$\text{METEOR} = (1 - \text{Pen}) \cdot \frac{PR}{\alpha P + (1-\alpha)R}, \tag{8}$$

where P and R are unigram precision and recall after alignment.

ChrF is the character n-gram F-score:

$$\text{ChrF}_\beta = \frac{(1+\beta^2)P_{\text{char}}R_{\text{char}}}{\beta^2 P_{\text{char}} + R_{\text{char}}}, \tag{9}$$

where P_{char} and R_{char} are averaged character n-gram precision and recall.

ROUGE-n is recall over word n-grams:

$$\text{ROUGE-}n = \frac{\sum_{g \in G_n(y)} \min\left(\text{count}_{\hat{y}}(g), \text{count}_y(g)\right)}{\sum_{g \in G_n(y)} \text{count}_y(g)}. \tag{10}$$

ROUGE-L is based on the longest common subsequence (LCS):

$$R_{\text{LCS}} = \frac{\text{LCS}(\hat{y}, y)}{|y|}, \qquad P_{\text{LCS}} = \frac{\text{LCS}(\hat{y}, y)}{|\hat{y}|}, \tag{11}$$

$$\text{ROUGE-}L = \frac{(1+\beta^2)P_{\text{LCS}}R_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}. \tag{12}$$

Lexical metrics are useful because the evaluation includes natural-language and short-answer QA in addition to structured extraction. For a strict field extraction, normalized exact match, precision, recall, F1, and hallucination rate are more appropriate; we report those metrics separately when the output is a structured ontology-keyed JSON object.

Model	Type	BLEU	METEOR	ChrF	R-1	R-2	R-L
ZS-2B	Extractive	15.9	62.2	71.8	78.8	44.5	78.8
	Boolean	0.0	0.8	11.5	53.3	0.0	53.3
	Grounding	0.0	0.0	1.1	0.0	0.0	0.0
LoRA FT-2B	Extractive	100.0	79.8	100.0	100.0	61.2	100.0
	Boolean	100.0	50.0	100.0	100.0	0.0	100.0
	Grounding	42.2	36.1	60.8	70.6	41.7	70.6

Table 12: Grounded QA performance by question type for Qwen3.5-2B. R-1, R-2, and R-L denote ROUGE-1, ROUGE-2, and ROUGE-L.

Metrics by Question Type Table 12 breaks down the metrics by question type. Finetuning yields improved performance on the extractive question answering, indicating that the generated corpus provides effective supervision for visible field extraction and field-level grounding.

B.3 Ontology-grounded extraction

Data Preparation Table 13 compares controlled synthetic training-data ablations for the same held-out ontology-keyed JSON extraction task.

OntoForm. Full ontology-grounded data with coherent entities, aligned schema bindings, rendered values, bounding boxes, and cross-form constraints.

Field-value mismatch. Data where the structured payload and visible rendered field values are intentionally misaligned.

Training data	Exact	Precision	Recall	$\mathbf{F_1}$
Field-value mismatch	40.90	98.30	77.40	86.60
Geometric corruption	41.20	98.10	78.00	86.90
Random field filling	42.20	95.50	79.90	87.00
ONTOFORM	**51.31**	**98.90**	**86.61**	**92.35**

Table 13: Ontology-keyed extraction after training on controlled synthetic variants (percent). "Exact" denotes exact-value accuracy over the evaluated fields and excludes correctly omitted null fields. The ONTOFORM row uses ontology-guided fine-tuning with ontology-conditioned inference.

Geometric corruption. Data where layout or geometry-token grounding is corrupted, while much of the text and payload content is preserved.

RandomFill. A non-ontology baseline with random type-valid field values and no enforced semantic or cross-form consistency.

Controlled Unseen-Form and Unseen-LOB Transfer The unseen-LOB evaluation uses 200 matched business-entity profiles shared across the two target LOBs. Each profile is rendered once per LOB as a four-page form, yielding 200 forms and 800 pages for Workers' Compensation and the same numbers for Commercial Property. The Workers' Compensation forms contain 171 ontology fields each, producing 34,200 evaluated field instances, of which 28,432 are populated. The Commercial Property forms contain 194 ontology fields each, producing 38,800 evaluated field instances, of which 34,456 are populated. In total, the evaluation covers 400 forms, 1,600 pages, and 73,000 field instances across 334 distinct ontology paths. Evaluated field counts include both populated and null fields.

The matched ontology configuration achieves the highest canonical path-value F_1 on both unseen LOBs. On Workers' Compensation, it improves canonical F_1 from 4.1% to 91.4% and reduces hallucination from 95.1% to 1.0%. On Commercial Property, it improves canonical F_1 from 5.2% to 80.2% and reduces hallucination from 93.5% to 0.1%. Hallucination rate is the proportion of gold-null fields for which the model predicts a non-null value.

Metrics We report parse rate separately and compute field-level metrics after type-aware normalization. Correctly populated values are true positives; hallucinated non-null values are false positives; missed rendered values are false negatives; and correctly omitted absent fields are true negatives.

Let $\mathcal{F}$ be the set of ontology fields evaluated for a document, and let y_i and $\hat{y}_i$ denote the normalized gold and predicted values for field $i \in \mathcal{F}$. We define $\mathrm{pop}(x) = 1$ if value x is non-empty after type-aware normalization, and 0 otherwise. Exact value match is denoted by $\mathbf{1}[\hat{y}_i = y_i]$.

Exact-match accuracy. Exact-match accuracy measures correctly recovered populated values over all evaluated ontology fields, without giving credit for correctly omitted null fields:

$$\mathrm{Exact} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \mathbf{1}[\mathrm{pop}(y_i) = 1]\, \mathbf{1}[\hat{y}_i = y_i]. \tag{13}$$

Accuracy including true negatives. To additionally credit correctly omitted null fields, we compute

$$\mathrm{Acc{+}TN} = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} \mathbf{1}[\hat{y}_i = y_i], \tag{14}$$

where $\hat{y}_i = y_i = \varnothing$ is counted as a correct true-negative prediction.

Precision. Measures how many populated predictions are correct:

$$\mathrm{Prec} = \frac{\sum_{i \in \mathcal{F}} \mathbf{1}[\mathrm{pop}(\hat{y}_i) = 1]\mathbf{1}[\hat{y}_i = y_i]}{\sum_{i \in \mathcal{F}} \mathbf{1}[\mathrm{pop}(\hat{y}_i) = 1]}. \tag{15}$$

Recall. Measures how many populated gold fields are recovered:

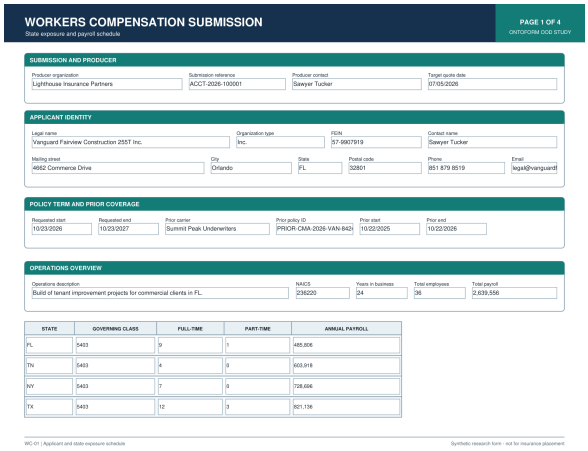
$$\mathrm{Recall} = \frac{\sum_{i \in \mathcal{F}} \mathbf{1}[\mathrm{pop}(y_i) = 1]\mathbf{1}[\hat{y}_i = y_i]}{\sum_{i \in \mathcal{F}} \mathbf{1}[\mathrm{pop}(y_i) = 1]}. \tag{16}$$

F1. The harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Prec} \cdot \mathrm{Recall}}{\mathrm{Prec} + \mathrm{Recall}}. \tag{17}$$

All comparisons are performed after type-aware normalization, including normalization of dates, numbers, currency values, checkbox states, and empty values.

(a) Commercial Property



(b) Workers' Compensation

Figure 5: Examples of the three digitally rendered surrogate insurance form types used in unseen-LOB transfer

FT/Condition	Canonical $F_1 \uparrow$	Key-agnostic $F_1 \uparrow$	Schema valid ↑	Correct null ↑	Hallucination ↓
Workers' Compensation					
No-Ont./No-Ont.	4.1±1.2	52.6±0.7	0.0±0.0	0.0±0.0	95.1±1.3
No-Ont./Ont.	5.3±9.2	74.1±20.7	5.5±9.6	0.0±0.0	94.5±9.5
Ont./No-Ont.	0.0±0.0	61.0±2.2	0.0±0.0	0.0±0.0	100.0±0.0
Ont./Ont.	**91.4±1.5**	**92.2±0.7**	**80.9±20.3**	**89.5±12.0**	**1.0±1.1**
Commercial Property					
No-Ont./No-Ont.	5.2±4.4	47.8±5.2	0.0±0.0	35.0±26.3	93.5±5.4
No-Ont./Ont.	1.1±2.0	86.7±4.5	1.8±3.2	0.0±0.0	98.9±1.9
Ont./No-Ont.	0.0±0.0	62.3±3.5	0.0±0.0	0.0±0.0	100.0±0.0
Ont./Ont.	**80.2±14.8**	**94.7±1.3**	**88.6±10.1**	**96.3±4.9**	**0.1±0.1**

Table 14: Complete controlled unseen-LOB transfer evaluation. All values are percentages. Higher is better for canonical F_1, key-agnostic F_1, schema validity, and correct-null rate; lower is better for hallucination rate.