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

EMNLP2026 Budapest, Hungary

Openstream.ai® Delivering Intelligence

Why ONTOFORM?

Challenge: Enterprise forms are linked views of applicants, locations, workforce and policies shared across a *submission packet*. We need training data that is both **available** and **consistent**, which requires two separate design choices.

A Why synthetic forms?

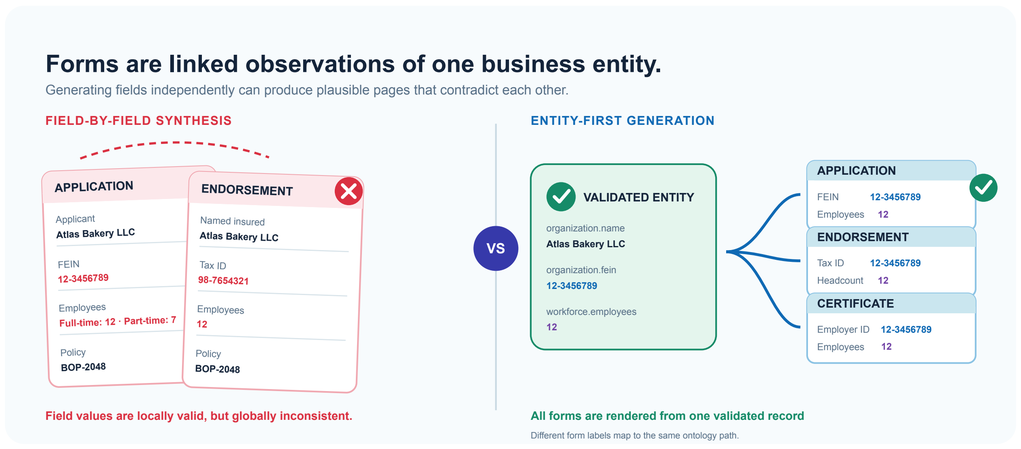
- Authentic documents are **privacy-restricted**.
- Box, field and QA labels are **expensive** to annotate by hand.
- Public benchmarks label pages **in isolation**: no linked multi-form packets.

⇒ Generate documents *with* exact labels.

B Why ontology-grounded?

- Filling field-by-field gives pages that are **each plausible, jointly impossible**.
- One concept, many labels: *Applicant* = *Named insured*.
- Stable semantic targets let models **transfer to unseen forms** and absent fields.

⇒ Generate the *entity* semantically with valid attributes



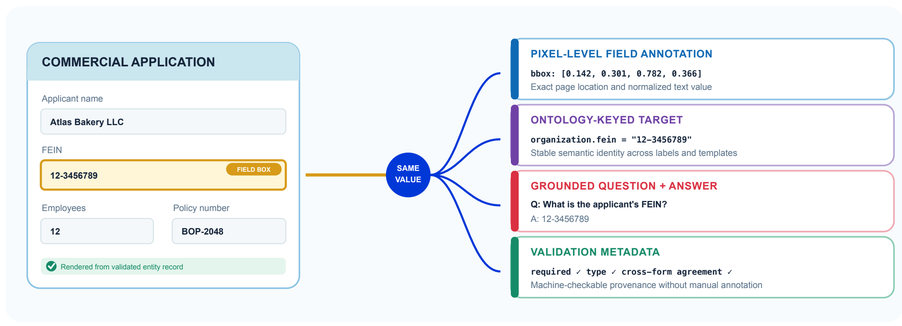
What is ONTOFORM?

ONTOFORM is a **document compiler** driven by a task-specific ontology: it turns one latent business entity into a consistent *packet* of rendered forms, each with exact training targets.

Approach: sample a validated entity record from the ontology, then render it into multiple form templates.

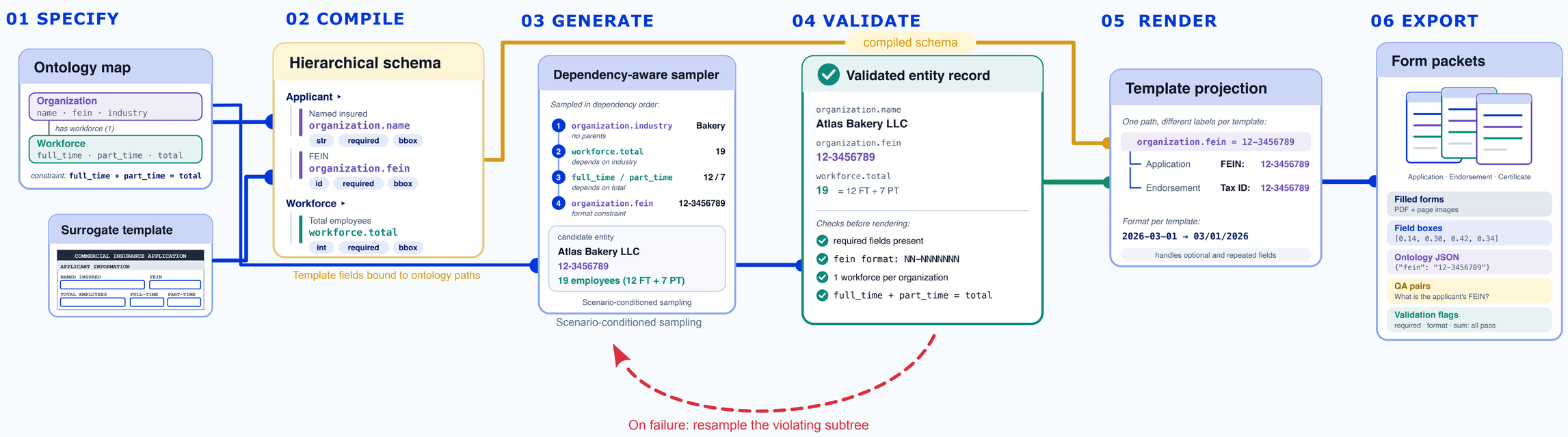
Core components

- **Ontology map:** concepts, value types, cardinalities, and dependencies.
- **Hierarchical schema:** template fields bound to stable ontology paths.
- **Entity sampler:** dependency-aware generation of coherent profiles.
- **Validator + renderer:** deterministic checks before projection.



The ONTOFORM Framework

Semantic-first generation: validate one entity, then render it into many forms



MANUAL ontology map + template bindings **AUTOMATIC** entity records + form packets + labels

One validated semantic record anchors every rendered form and every supervision target.
Instantiated for insurance underwriting: **Commercial Application**, **General Liability** and **Business Auto** forms.

RESULTS

Ontology-Guided Extraction: Ontology supervision improves semantic alignment and omission handling.

Training	Ontology	Exact	Recall	Acc.+TN	F_1
Donut	×	75.65	75.65	–	85.68
Zero-shot	×	2.87	4.84	2.89	9.21
Zero-shot	✓	**54.11**	**91.33**	83.64	87.36
FT	×	47.53	80.22	75.17	88.63
FT	✓	51.31	86.61	**90.40**	**92.35**

Same model and budget; only ontology supervision changes: **+15.23 Acc.+TN** and **+3.72** F_1.

+44.1 **Grounded QA** 53.1 ⟶ **97.2** Grounded answers about a single rendered form: *"What is the applicant's FEIN?"*

Generalization Beyond Seen Forms: Train on Commercial + GL; test on unseen Business Auto concepts.

Scope	F_1 w/o $\mathcal{O}$	F_1 w/ $\mathcal{O}$	ΔF_1
All fields	60.90	**96.95**	**+36.05**
Shared concepts	93.45	**99.15**	**+5.70**
Held-out concepts	53.92	**96.59**	**+42.67**

The largest improvement occurs on concepts never observed during training, indicating semantic transfer rather than template memorization.

+34.17 **Cross-form reasoning** 18.33 ⟶ **52.50** Evidence combined across related forms in one packet: *"Do the Application and Certificate report the same headcount?"*