Formal Ontologies for LLM-Agent Knowledge Management: schema.org vs gUFO vs DOLCE

Jesse Wright · the sparq project

*DRAFT — a controlled **pilot** study. Every accuracy figure in this paper is an* indicative *measurement: the outcome of one fully-crossed run of a non-deterministic LLM agent (a small model, Claude Haiku), graded by a deterministic heuristic grader, on a development machine. No accuracy figure is canonical evidence, none is co-tabulated with canonical evidence, and no statistical significance is claimed. The pilot's effect directions motivate — but do not yet license — a venue submission; the multi-run study this draft commits to (Section 6.), to be pre-registered before it runs, is the gate.*

Abstract

An LLM agent that manages durable project knowledge — decisions, provenance, task structure, design records — increasingly does so over a *project knowledge graph* (PKG) that the agent itself queries. Knowledge engineering doctrine says such a graph should be typed under a formal upper ontology, but which one, and whether typing helps an LLM consumer *at all*, has not been measured. We compare four typing conditions over the same PKG instance data — no formal ontology, schema.org-as-top, gUFO, and DOLCE — on a corpus of 16 repository knowledge-management questions answered through SPARQL by a deliberately small LLM agent (Claude Haiku; one fresh agent instance per condition–task pair), in a single fully-crossed pilot run with deterministic heuristic grading. The indicative result: *two* overlays beat the untyped baseline's 0.58 accuracy — schema.org by the largest margin (0.84, and the top condition on every task kind) and DOLCE by a small one (0.64) — while gUFO fell slightly below it (0.54). The overlay ranking (schema.org > DOLCE > gUFO) is consistent with the vocabularies' plausible frequency in LLM training data, and the one overlay that hurt is the least web-prevalent; but DOLCE's gain shows a formally rich ontology can also convert into accuracy, and the pilot's design cannot separate vocabulary fluency from a competing mechanism (overlay size and closure noise). The supported reading is deliberately narrow: for this small agent model, on this FO-exercising corpus, in one run, the most fluent vocabulary helped most and the least fluent formal overlay was worse than none. This is a hypothesis-generating pilot, not a demonstrated mechanism.

1. Introduction

Agentic LLM systems accumulate knowledge they must later retrieve precisely: which design was chosen and why, what a measurement showed, which tasks block which. A natural substrate is a *project knowledge graph* — an RDF graph ingested from the project's durable knowledge surfaces (agent instructions, skill documents, the dependency-aware task tracker) that the agent queries with SPARQL instead of re-reading whole documents. This is the classical knowledge-management stack with a new consumer: the reader of the ontology is no longer a human or a description-logic reasoner but a large language model that must *translate its own natural-language framing of a question into the graph's vocabulary*.

Knowledge-engineering doctrine holds that a knowledge graph gains from a formal upper ontology: a top level such as DOLCE or UFO disambiguates categories (endurant vs perdurant, kind vs role), supports consistency checking, and makes modelling decisions explicit. But every argument in that tradition assumes a consumer that benefits from formal discipline. An LLM consumer is different in exactly the dimension the tradition ignores: its competence over a vocabulary is a function of how often that vocabulary occurs in its training distribution. schema.org terms are embedded in billions of web pages; DOLCE and UFO terms live in a specialist literature. Whether formal richness or distributional fluency dominates is an empirical question, and to our knowledge nobody had measured it.

This paper reports a controlled pilot that measures it. We hold the PKG instance data, the agent, the question corpus, and the grading fixed, and vary ONLY the typing overlay: (1) the untyped incumbent graph, (2) schema.org-as-top, (3) gUFO (the lightweight OWL distillation of UFO), and (4) DOLCE (via its DUL OWL rendering). A fresh instance of a small LLM agent (Claude Haiku) answers each of 16 knowledge-management questions per condition by querying the graph; answers are graded against gold facts by a deterministic heuristic grader.

The pilot's indicative finding is more textured than the fluency-versus-doctrine slogan we set out to test. *Two* overlays beat the untyped baseline: schema.org-as-top — the formally weakest vocabulary — by the largest margin, and DOLCE — a formally rich one — by a small margin. One overlay, gUFO, landed below the baseline. If this pattern survives the scale-up study (Section 6.), the practical guidance is conditional rather than absolute: typing a knowledge graph for an LLM agent can help or hurt depending on *which* vocabulary is chosen — the vocabulary the model plausibly knows best gained the most here, and the least web-prevalent one cost accuracy relative to doing nothing. Measure before you type, because typing is not free.

Contributions. This draft substantiates the following; each claim is refutable and forward-references its evidence. No accuracy claim is canonical, and none is a performance (latency/throughput) claim.

- **A controlled four-condition study design for ontology choice under an LLM consumer** (Section 3.) — same instance graph, same agent, same 16-question corpus, same grader; only the committed typing overlay (4 conditions) varies, fully crossed (every condition answers every task). The corpus, overlays, task builder, validator, and grader are committed and re-runnable (`bench/fokm/`), making the design a reusable resource independent of the pilot's numbers.
- **An indicative pilot result** (Section 4.) — in one fully-crossed run, two overlays improved over the untyped baseline: schema.org (0.84 vs 0.58) by far the most, DOLCE (0.64) modestly — and gUFO (0.54) fell below it. Explicitly indicative: single run, heuristic grading, non-deterministic agent, one small model, one date.
- **A fluency hypothesis with its confounds named** (Section 5.) — the overlay accuracy ranking is consistent with the vocabularies' plausible training-data frequency; we state fluency as a *correlational hypothesis* (the pilot's only mechanism evidence is a single ranking of point estimates against a hand-asserted frequency ordering), enumerate the competing overlay-verbosity/closure-noise mechanism the design cannot exclude, and specify the discriminating experiments that would separate them.
- **An honest evidence protocol for single-run agent-produced results** (Section 3.4., Section 6.) — every accuracy figure is routed through an evidence file that labels it `indicative` and a build gate that structurally prevents it from ever being cited as canonical headline evidence; the pre-submission obligations are enumerated on the face of the paper.

2. Related work

Upper ontologies. DOLCE (Masolo et al. 2003; Borgo et al. 2022) and UFO (Guizzardi 2005; Guizzardi et al. 2022) are the two foundational ontologies with the deepest methodological literature; gUFO (Almeida et al. 2019–) is UFO's lightweight OWL 2 DL rendering intended exactly for typing knowledge graphs, and DOLCE's DUL rendering plays the same role. BFO (Arp, Smith & Spear 2015) and SUMO (Niles & Pease 2001) serve adjacent niches. The selection literature in this tradition compares ontologies by formal criteria — categorial coverage, axiomatic rigour, modelling guidance — always assuming a human modeller or a DL reasoner as the consumer. We are not aware of prior work that selects an upper ontology by measured downstream task accuracy of an LLM consumer; this is the gap the pilot addresses.

schema.org. schema.org (Guha, Brickley & Macbeth 2016) was designed with the opposite priorities: a shallow, wide, deliberately informal vocabulary optimised for mass adoption by web authors and consumption by search engines. Its formal weaknesses (thin axiomatization, permissive domains/ranges) are documented and real. Precisely because of its adoption, however, its terms are among the highest-frequency structured-data vocabulary in any web-scale training corpus — the property our mechanism hypothesis (Section 5.) turns on.

LLMs and knowledge graphs. The LLM+KG literature (Pan et al. 2024) spans KG-augmented prompting (Baek et al. 2023), iterative graph reasoning (Sun et al. 2024), and graph-indexed retrieval (Edge et al. 2024). Closest to us, Sequeda, Allemang & Jacob (2024) and Allemang & Sequeda (2024) show that putting an ontology (and a SPARQL-over-ontology layer) between an LLM and enterprise data substantially improves question-answering accuracy over schema-less access. Their comparison is *ontology vs no ontology* for one fixed ontology; ours holds "there is an ontology" fixed and varies *which* ontology, finding the choice can swing the outcome from a gain to a regression. Work on LLMs as ontology engineers (Babaei Giglou et al. 2023; Caufield et al. 2024) measures the converse direction (LLMs producing ontology artifacts) and does not compare upper ontologies as consumers' vocabularies. Agent-memory systems (Packer et al. 2023; Park et al. 2023) and personal knowledge graphs (Balog & Kenter 2019) motivate the PKG substrate but do not type it formally.

Positioning. The nearest prior art establishes that structure helps LLM question answering (Sequeda, Allemang & Jacob 2024; Allemang & Sequeda 2024). Our delta is the controlled comparison *across* upper-ontology choices with an untyped control, and the sign-sensitivity of the finding — one formal overlay (gUFO) landed *below* the untyped baseline while another (DOLCE) landed above it, i.e. the choice of upper ontology can flip the sign of the intervention — which we have not found reported elsewhere.

3. Study design

3.1. Substrate: a project knowledge graph with a real workload

The substrate is the PKG of a working RDF-engine repository: an RDF graph ingested from the project's durable knowledge surfaces — the agent-instructions document, the per-surface skill documents, and the dependency-aware task tracker. The graph is in production use by the project's own agents, which answer provenance and status questions over it via SPARQL rather than re-reading documents. The workload is therefore not synthetic: the question corpus (Section 3.3.) is drawn from the question shapes those agents actually pose ("where was X decided", "what is the status and provenance of Y", "what depends on task Z").

3.2. Conditions: four committed typing overlays over one instance graph

The experimental variable is a *typing overlay*: a Turtle document asserting class membership (and, where the ontology provides them, object-property specialisations) for the PKG's instance data under one top-level vocabulary. 4 overlays are committed under `bench/fo-km/overlays/`:

- **no-FO** (`no-fo.ttl`) — the incumbent untyped control: the PKG as ingested, with only its native ad-hoc vocabulary.
- **schema.org** (`schema-org.ttl`) — instances typed under schema.org classes (e.g. `schema:SoftwareSourceCode`, `schema:CreativeWork`, `schema:Action`) with schema.org properties layered over the native predicates.
- **gUFO** (`gufo.ttl`) — instances typed under gUFO's OWL rendering of UFO categories (kinds, phases, roles, relators, situations).
- **DOLCE** (`dolce-dul.ttl`) — instances typed under the DOLCE+DnS Ultralite (DUL) rendering (endurant/perdurant/quality/abstract partitions and DUL's description-and-situation pattern).

The instance data is identical across conditions; only the overlay differs. Each condition is materialised as the instance graph plus its overlay loaded under an OWL-RL closure, so the overlay's subclass axioms entail the FO-typed facts the tasks exercise — a detail that matters for the mechanism discussion (Section 5.), because richer overlays entail larger closed graphs. Overlay authoring followed the target ontology's own modelling guidance.

3.3. Task corpus and run protocol

The corpus is 16 knowledge-management questions (`bench/fo-km/tasks.jsonl`, one task per line, schema-validated by `bench/fo-km/validate_tasks.py`), each pairing a natural-language question about the repository's knowledge estate with gold facts an answer must contain, stratified over three task kinds (type-hierarchy, entailment-dependent, and cross-category questions). The corpus is FO-exercising *by construction*: every task is validated to discriminate between conditions — a typed condition can answer it under closure while the untyped graph returns no rows and must hand-enumerate or abstain — and questions answerable by the graph's native vocabulary alone were deliberately excluded. This is a scope decision the reader must carry throughout: the corpus measures what typing buys *where typing can matter at all*, not the agent's average workload (Section 6.).

The design is fully crossed: every condition answers every task (4 × 16 condition–task pairs), with one fresh agent instance per pair, so no condition sees a different question set and no instance carries state between tasks. The agent is a deliberately small LLM (Claude Haiku). Each instance receives only the natural-language question and its condition's overlay; it answers end to end by introspecting the graph's vocabulary, composing SPARQL against the overlaid PKG, and synthesising an answer from the bindings. It never sees the gold facts or any reference query. The pilot consists of a *single* such pass — a deliberate scope choice for a pilot, and the study's primary limitation (Section 6.).

3.4. Grading, abstention, and the evidence protocol

Answers are graded by a deterministic heuristic grader (`bench/fo-km/analyze.py`) with no model in the grading loop: a task is graded correct when the answer covers its gold keys — a count/list task must resolve the gold count and entity names, and a partition task must cover each gold sub-category. Deterministic grading removes grader non-determinism, but not grader bias: coverage heuristics can over-credit keyword overlap and under-credit paraphrase (Section 6.).

An agent instance that concludes its condition genuinely cannot answer a task, and says so, is scored as an *abstention* — counted separately, not as a wrong answer. Abstention rates differ sharply across conditions (from 1 to 5 of 16 tasks; Section 4.), which is itself a confound: the four accuracies do not rest on identical answered-task mixes. The per-condition accuracy we report is the measurement record's task-kind-weighted graded-correct fraction of the corpus.

Every accuracy figure is routed through the project's evidence file, where each record carries `environment: "indicative"`, the run provenance, and the grader source. The build machinery structurally refuses to render an indicative record as headline evidence — the same gate that keeps non-canonical timing out of the project's other papers keeps single-run agent-produced accuracy out of any headline here. The only canonical records this paper cites are the two structural corpus counts (16 tasks, 4 conditions).

evidence: `bench/fo-km/tasks.jsonl (one task per line; schema checked by bench/fo-km/validate_tasks.py)` (environment: canonical)

4. Pilot results (indicative)

Condition	Answer accuracy	Abstained	Relation to untyped baseline
schema.org-as-top	0.84	2/16	above (largest margin)
DOLCE (DUL overlay)	0.64	1/16	above
no-FO (untyped control)	0.58	5/16	—
gUFO	0.54	2/16	below

Table 1: INDICATIVE pilot measurements — not canonical evidence. Task-kind-weighted graded-correct fraction (and abstention count) over the 16-task corpus in ONE fully-crossed run of a non-deterministic LLM agent (Claude Haiku) with deterministic heuristic grading on a development machine (`environment: indicative` in the project's evidence file; see Section 3.4.). Abstentions are scored separately, not as wrong answers (Section 3.4.). No significance test is possible or claimed at this scale. Values transcribed directly from the measurement record `bench/fo-km/RESULTS.md`.

Four readings, each stated at pilot strength:

The most web-prevalent vocabulary gained the most. schema.org's 0.84 against the baseline's 0.58 is an absolute gap of 0.26 — roughly 4 additional tasks of 16 — and, per the measurement record, schema.org was the top condition on every one of the three task kinds. At this N the gap is directionally interesting and nothing more; it is the effect the scale-up study must confirm or kill.

A formally rich overlay also beat the baseline. DOLCE's 0.64 against 0.58 is a small gain — on the order of one task — and its direction matters more than its size: it contradicts any strong "formal richness cannot convert into LLM-agent accuracy" reading. This is why the paper states fluency-over-formality as a hypothesis about *magnitudes* (the most fluent vocabulary gained most) rather than a law (only the fluent vocabulary gains). The record also shows DOLCE abstaining least (1/16): its method/document/description categories gave the agent reachable targets.

One overlay was worse than nothing. gUFO's 0.54 falls below the untyped baseline by roughly one task — small, and single-run — but the direction is the practically alarming one: typing the graph is not a monotone improvement. An overlay whose vocabulary the agent handles poorly displaces native terms the agent handled adequately, and the net effect can be negative.

Abstention differed five-fold across conditions. The untyped control abstained on 5 of 16 tasks against 1 for DOLCE (Section 3.4.). This is partly by construction — the corpus deliberately contains tasks the untyped graph cannot answer (Section 3.3.) — but it also means the four accuracies ride on different answered-task mixes: an imbalance the reader must weigh, and one the scale-up study must control (for example by additionally reporting accuracy over the subset of tasks every condition answered).

5. A fluency hypothesis — and the confounds the pilot cannot exclude

What the pilot licenses is exactly this: a ranking of four point estimates from one run, in which the overlay ordering (schema.org > DOLCE > gUFO) is consistent with a plausible — asserted, not measured — ordering of the vocabularies' frequency in web-scale training corpora. schema.org is embedded in billions of pages as JSON-LD and microdata; DOLCE/DUL has two decades of academic and linked-data usage; gUFO, the most recent and most specialised of the three, is plausibly the rarest. No direct fluency measurement was taken; the frequency ordering is the authors' judgment; and the untyped baseline's position between DOLCE and gUFO is co-determined by the corpus construction (the corpus contains tasks the untyped graph cannot answer; Section 3.3.). This section's heading names a hypothesis, not a result.

The fluency mechanism. To answer a KM question the agent must map its natural-language framing onto graph terms. A high-fluency vocabulary acts as a set of familiar anchors: `schema:Action`, `schema:about`, `schema:isPartOf` mean to the model roughly what they mean in the graph, so the composed SPARQL binds the right classes on the first attempt. A low-fluency vocabulary (gUFO kinds/phases/relators; DOLCE endurants/perdurants) inserts an unfamiliar indirection layer between the question and the data: the agent must first infer what the ontology means by its terms — a step where a rare technical vocabulary invites exactly the near-miss (a plausible-but-wrong class pick) that grades as a wrong answer.

The competing mechanism the design cannot exclude. The overlays do not only change vocabulary; they change the *size and noise of the graph the agent works in.* A richer overlay under OWL-RL closure (Section 3.) yields a larger entailed graph and a bigger introspection surface — more classes, more entailed rows, longer contexts for a small model to wade through, independently of whether it knows the terms. The project's own adjacent experience points the same way: in a separate internal experiment over the same knowledge graph, loading a formal overlay with alignment axioms under closure produced an entailment-noisy graph and bloated introspection output that measurably degraded a cheap-model consumer, independent of vocabulary choice. The pilot's record offers only a partial internal check: by its token accounting, DOLCE drove the *largest* introspection surface of the four conditions yet gained accuracy, while gUFO drove the *smallest* yet lost — so surface size alone does not order the outcome either. But nothing in this design isolates vocabulary from verbosity, and attributing the whole effect to fluency — as an earlier draft of this paper did — is not supported.

Discriminating experiments (all future work; none has been run). The hypothesis is falsifiable in at least four ways the scale-up study can operationalise: (a) a per-error analysis attributing failures to vocabulary-translation misses vs retrieval or synthesis misses; (b) a fluency probe (does the model define/instantiate each vocabulary term correctly in isolation?) whose per-term scores should predict per-task outcomes if the mechanism is real; (c) an ablation that renames gUFO/DOLCE classes to descriptive plain-English aliases while keeping the class *structure* — if fluency is the driver, aliasing should recover most of the gap; if formal structure is the driver, it should not; and (d) a verbosity/closure ablation that holds the vocabulary fixed while varying the entailed-graph size (closure on vs off; truncated overlays) — if context noise is the driver, equalising the surface should erase the vocabulary effect.

Model-relativity. The hypothesis is explicitly model-relative and time-indexed: fluency is a property of a training distribution, not of an ontology — and the pilot's agent is a deliberately small, cheap model (Claude Haiku). A fluency deficit is plausibly *amplified* at that scale; a frontier model may wield `gufo:` and `dul:` terms without difficulty, which would shrink or erase the effect. Nothing in this pilot licenses a claim about "LLM consumers" in general. The finding must be reported per model and date, and the scale-up study must span model tiers.

6. Limitations and threats to validity

This section is the paper's honesty boundary; the draft-status box below restates the pre-submission obligations operationally.

- **Pilot scale.** 16 tasks, one fully-crossed run. The headline gap is roughly four tasks; the DOLCE and gUFO deltas are on the order of *one* task each — well within plausible single-run variance of a non-deterministic agent. No confidence intervals or significance tests are possible, and none are claimed. This scale is disqualifying for a top-tier research track on its own; it is why the paper is a pilot.
- **Heuristic grading.** The grader is deterministic (no model in the grading loop) but coverage-heuristic: it can both over-credit keyword overlap and under-credit paraphrase. Grading error is not guaranteed to be condition-independent (conditions change the agent's phrasing), which is a bias

channel, not just noise. The scale-up study needs human-adjudicated (or at minimum adversarially-validated) grading.

- **Abstention asymmetry.** Abstention rates spread five-fold across conditions (5 vs 1 of 16 Section 3.4.), so the accuracies compare different answered-task mixes; and the exact denominator arithmetic for abstentions lives in the grader source and is a draft obligation to restate (Section 3.4.).
- **Corpus construction.** The corpus is FO-exercising by construction (Section 3.3.): tasks answerable by the native vocabulary alone were excluded, and some tasks the untyped graph cannot answer at all. The comparison therefore measures typing where typing can matter; on a mixed, average workload the deltas would dilute — and could change sign.
- **One small model, one date.** The agent is Claude Haiku, a small/cheap model tier, run once at one point in time. The fluency deficit is plausibly amplified at that tier (Section 5.); generalisation across model families and tiers is untested, and the paper makes no claim about LLM consumers in general.
- **One domain.** The PKG is a software-project knowledge graph; KM over other domains (scientific, enterprise) may reward ontological precision differently.
- **Overlay-quality asymmetry — the deepest open threat.** The overlays were hand-authored by authors who are themselves more fluent in schema.org than in gUFO or DOLCE, so the measured ordering could be an artifact of better-authored schema.org overlays rather than of the model's fluency. The countermeasures — independent expert review of the gUFO/DOLCE overlays and the aliasing ablation (Section 5., experiment (c)) — have *not been run*. Until they are, the fluency reading cannot be asserted as a conclusion; it remains a hypothesis contaminated by an unclosed confound, and the paper says so wherever the hypothesis appears.
- **Mechanism not isolated.** The design confounds vocabulary choice with overlay verbosity and closure noise (Section 5.); no mechanism-discriminating experiment has been run.
- **Deferred second metric.** The pilot's planned graph-representational metric (a knowledge-graph-embedding evaluation over the same overlays) was deferred for compute and is not reported; the accuracy metric stands alone.
- **No pre-registration of the scale-up.** The pilot's harness design record registered a neutral prior before the run — the only registration artifact that exists. The multi-run scale-up this paper commits to is *not yet pre-registered*; "to be pre-registered" is a commitment about future work, never a present rigor credential.
- **Non-canonical environment.** All measurements executed on a development work-box; per the project's standing evidence policy such numbers are indicative and are labelled so throughout.

Draft status — obligations before any venue submission (tracked as bead `sq-iw378`): (1) ~~verify all accuracy figures against `bench/fo-km/RESULTS.md`~~ — done in this revision: all four accuracies and abstention counts are transcribed directly from the measurement record (the first draft's mis-transcribed baseline and fabricated two-figure gUFO range are corrected); (2) pin the exact agent model version string, run date, and prompting scaffold from the run records, and restate the grader's abstention/denominator arithmetic from `analyze.py`; (3) recover and cite the overlay-construction protocol; commission independent review of the gUFO/DOLCE overlays; (4) write and register the scale-up pre-registration *before* it runs, then execute the multi-run study with human-adjudicated grading, confidence intervals, and the discriminating experiments of Section 5. (per-error attribution, fluency probe, aliasing ablation, verbosity/closure ablation), across model tiers; (5) run the deferred KGE metric; (6) pin bibliography DOIs/pages (author/venue/year were checked against public records in this revision); (7) double-blind readiness: build with the `anon` toggle (strips the author block and the non-anon footer) *and* de-identify the substrate description (Section 3.) for any double-blind venue.

7. Conclusion

We measured a question knowledge-engineering doctrine had answered only by argument: whether a formal upper ontology helps when the consumer of the knowledge graph is an LLM agent, and which one. In a fully-crossed four-condition pilot over a production project knowledge graph (16 tasks, 4 committed overlay conditions, one small-model agent), the indicative answer was neither the doctrine's ordering nor its clean inversion. Two overlays beat the untyped baseline: schema.org-as-top — the formally weakest, most web-prevalent vocabulary — by the largest margin (0.84 vs 0.58, and top on every task kind), and DOLCE — a formally rich vocabulary — by a small one (0.64). One overlay, gUFO, landed below the baseline (0.54). The overlay ranking is consistent with training-data fluency; DOLCE's gain shows formal richness can also convert into accuracy; and the pilot cannot separate fluency from overlay verbosity and closure noise as the driver — so we state the mechanism as a falsifiable hypothesis with the ablations that would decide it. The pilot is honest about being a pilot: every accuracy figure is single-run, heuristically graded, produced by one small model on one date, environment-labelled indicative, and structurally barred from ever being cited as canonical evidence; the multi-run scale-up, to be pre-registered before it runs, is the gate to any venue. If the pattern survives, the practical rule for anyone typing a knowledge graph for an LLM agent is conditional and slightly uncomfortable: *which* ontology you choose can decide whether typing helps or hurts — here, the vocabulary the model already spoke gained the most, and the one it spoke least was worse than none. Measure before you type.

References — author/venue/year of the empirical LLM+KG entries were checked against public records during this revision (2026-07-01); foundational-ontology entries are standard citations. DOIs and page numbers remain to be pinned in the camera-ready bibliography (Draft-status obligation (6)).

Masolo et al. 2003 *WonderWeb Deliverable D18: Ontology Library* (DOLCE). Masolo, Borgo, Gangemi, Guarino, Oltramari.

Borgo et al. 2022 "DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering." Borgo, Ferrario, Gangemi, Guarino, Masolo, Porello, Sanfilippo, Vieu. *Applied Ontology* 17(1).

Guizzardi 2005 *Ontological Foundations for Structural Conceptual Models.* PhD thesis, University of Twente.

Guizzardi et al. 2022 "UFO: Unified Foundational Ontology." Guizzardi, Botti Benevides, Fonseca, Porello, Almeida, Prince Sales. *Applied Ontology* 17(1).

Almeida et al. 2019– *gUFO: A Lightweight Implementation of the Unified Foundational Ontology.* Almeida, Guizzardi, Falbo, Prince Sales. OWL vocabulary.

Arp, Smith & Spear 2015 *Building Ontologies with Basic Formal Ontology.* MIT Press.

Niles & Pease 2001 "Towards a Standard Upper Ontology." *FOIS.*

Guha, Brickley & Macbeth 2016 "Schema.org: Evolution of Structured Data on the Web." *CACM* 59(2).

Pan et al. 2024 "Unifying Large Language Models and Knowledge Graphs: A Roadmap." Pan, Luo, Wang, Chen, Wang, Wu. *IEEE TKDE.*

Baek et al. 2023 "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering." Baek, Aji, Saffari. *NLRSE workshop @ ACL 2023* (arXiv:2306.04136).

Sun et al. 2024 "Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph." *ICLR.*

Edge et al. 2024 "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." arXiv:2404.16130 (preprint).

Sequeda, Allemang & Jacob 2024 "A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases." *GRADES-NDA @ SIGMOD 2024* (arXiv:2311.07509, 2023).

Allemang & Sequeda 2024 "Increasing the LLM Accuracy for Question Answering: Ontologies to the Rescue!" arXiv:2405.11706 (preprint).

Babaei Giglou et al. 2023 "LLMs4OL: Large Language Models for Ontology Learning." Babaei Giglou, D'Souza, Auer. *ISWC.*

Caufield et al. 2024 "Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES)." *Bioinformatics*.
Packer et al. 2023 "MemGPT: Towards LLMs as Operating Systems." arXiv preprint.
Park et al. 2023 "Generative Agents: Interactive Simulacra of Human Behavior." *UIST*.
Balog & Kenter 2019 "Personal Knowledge Graphs: A Research Agenda." *ICTIR*.

---

sparq project · the study corpus, overlays, builder, validator, grader, and results record live under `bench/fo-km/` (`tasks.jsonl`, `overlays/`, `build_tasks.py`, `validate_tasks.py`, `analyze.py`, `RESULTS.md`). Numbers in this document are injected at build time from the paper-bound evidence file (accuracy and abstention records are `environment: indicative`; only the structural corpus counts are canonical); see the provenance stamp on the published page. Draft authored under the Fable provenance protocol from the audited prose digest (`research/papers-venue-audit.md`); revised against the measurement record `bench/fo-km/RESULTS.md` in the PR 1335 review.