

Peer Review

Review of: "Top Ten Challenges Towards Agentic Neural Graph Databases"

Evgenii Arsentev¹

1. Independent researcher

Summary

The paper proposes "Agentic Neural Graph Databases" (Agentic NGDBs) as the next step after Neural Graph Databases, defined by three functionalities — autonomous query construction, neural query execution, and continuous learning (Abstract; §1, final paragraph). It then organises ten challenges under three headings: Interface (Challenges 1–2), Learning and Inference (Challenges 3–5), and System (Challenges 6–10), and devotes one section to each (§2 through §11). This is a position/agenda paper, not an empirical one: it reports no experiment, no dataset of its own, and no table. Its load-bearing content is therefore the taxonomy itself, the literature it maps onto each challenge, and the argument that the ten items are the right ten. I read it on those terms. The mapping of Challenges 1–5 onto the complex logical query answering literature is genuinely useful and, in places (§4.1, §4.2, §6), better than what a reader would assemble from the surveys alone. The framing around it is weaker: the selection criterion for the ten is never stated, several sections state hoped-for properties in the indicative mood, one of the ten is not a challenge at all, and the functionality the title leans on most — autonomy, i.e. a system that issues its own CREATE/UPDATE/DELETE — is the one the challenge list covers least.

Does the claim hold on the data

There is no data, and that is fine for this genre, so the question becomes: is each claim carried by a citation, and is the list complete and non-arbitrary? Two things stand out.

First, the completeness argument is a single sentence. §1, the paragraph immediately before Figure 1: "We identify the most critical challenges for realizing these functionalities based on recent progress in the research community on logical query answering and logical hypothesis generation for relational graphs." That is the whole method. It also scopes the evidence base to two subfields, which happen to be the

authors' own; a reader cannot tell whether an eleventh challenge was considered and rejected, or never came up. For a paper whose title is a ranked-sounding list ("Top Ten"), that is the claim most in need of support, and it has the least.

Second, several passages describe a system that does not yet exist using verbs reserved for one that does. §10 (Challenge 9): "NGDB-RAG enhances retrieval by utilizing the structured relationships in NGDBs"; "NGDB-RAG integrates structured knowledge and reasoning capabilities from NGDBs to improve the generated text's factual accuracy and logical consistency while reducing hallucinations"; "NGDB-RAG is designed to handle large-scale graphs and supports various query types, including temporal, spatial, and numerical reasoning." §11 (Challenge 10): "Agentic NGDB can autonomously manage complex datasets, optimize query execution, and organize storage structures without human intervention." §8: "Such compatibility design will enable a seamless system transition." §4.2: "Notably, the neuro-symbolic approach achieves the same level of generalizability in queries as the classic database research." None of these is hedged, and none is measured. The fix is mostly grammatical — subjunctive instead of indicative, plus a pointer to what would have to be shown — but as written a hurried reader comes away believing NGDB-RAG has been built and evaluated.

On numbers: the paper contains almost none, so there is little to check. The exceptions are the complexity expressions in §6 ("Test Time Scaling in Inference Stage"): $O(|Q||E|)$ for tree-formed queries, $O(|E|^{|Q|})$ for general EFO-1, and $O(|Q|)$ encode plus $O(|E|)$ decode for query embeddings and sequence models; and the objective $L_{\text{total}} = L_{\text{LLM}} + \lambda L_{\text{NGDB}}$ in §10 ("Joint Learning Framework"). I raise concrete issues with both below. The one quantitative claim about results — §3, second paragraph, "achieving significant gains in metrics like Jaccard similarity and Smatch scores across multiple datasets" — carries no number, no dataset name at that point, no baseline, and no test, yet uses "significant."

A note required by how the article reached me: Figure 1 arrives as a caption only ("Figure 1. The top ten challenges in achieving Agentic NGDB. Its three perspectives include interface, learning, and system."), with no rendered content, and the article contains no tables. Several inline formulas in §4.1, §6, and §10 are duplicated in plain and mathematical form in the version I have (e.g., "(n+1)" then "(n + 1)"), which I take to be an extraction artefact rather than an authoring one. I have not judged anything on those grounds.

Reproducibility

No code, data, models, or runs are involved, so the usual checklist does not apply. The analogue for a challenge paper does, and it is the single change that would most raise this paper's value: *make each challenge falsifiable*. As it stands, nowhere in §2–§11 is there a statement of what would count as having solved, or measurably advanced, any of the ten. No reference task, no dataset, no metric, no baseline, no target. §5.2 ("Evaluation Benchmarks") comes closest — it correctly identifies that privacy benchmarks for NGDBs are missing — but then asks for them rather than specifying what they should measure. The only datasets named anywhere are FB15k-237 and WN18RR, and only in passing in §3 as what reference [15] trained on.

Concretely, I would ask for a short table, one row per challenge, with four columns: the task, the dataset or benchmark that exists today (or "none"), the metric, and the number that current best practice achieves. Where the answer is "none," say so — that emptiness is itself the finding, and it is far more actionable than prose. This would also discipline the list: a challenge for which the authors cannot name a measurement is a challenge that may not yet be well-posed, and that test would probably resolve the Challenge 10 problem below on its own.

Two smaller reproducibility-adjacent asks. (1) §6 discusses scaling of parameters, query data, and training cost, but no parameter count, dataset size, or training budget appears anywhere in the paper for any existing CQA model. Even one row of orders of magnitude would let a reader judge whether scaling is an open opportunity or a saturated one. (2) §10's joint objective has a free hyperparameter λ and no guidance, no reference, and no discussion of the gradient-magnitude mismatch between a next-token cross-entropy and a query-answering loss. As printed, it is a placeholder, presented in the position where a contribution would go.

Specific points

1. **The selection criterion for the ten is missing, and the granularity is uneven enough to suggest the list was assembled rather than derived** (§1, paragraph before Figure 1; compare §4 with §11). Challenge 3 gets three subsections with a taxonomy of query families, minimal-assumption methods, and a learning-theoretic framing. Challenge 10 gets three bullets. Challenge 6 internally contains four distinct problems — user interface design, query-oriented distributed storage, distributed graph computing and inference, elastic scalability (§7.2) — any one of which is larger

than Challenge 7. Either state the rule that makes these ten items commensurable (e.g. "each is a problem where no system currently achieves X"), or drop the counted framing and present three clusters with sub-problems. The paper would lose nothing and gain honesty.

2. **The defining functionality has no challenge of its own: autonomous writes are proposed but their correctness is never posed as a problem.** §1, System bullet: the Agentic NGDB "must actively learn by constructing and executing appropriate CREATE, UPDATE, or DELETE queries in a given context." A database that deletes rows on its own initiative raises transactions, isolation, durability, rollback, provenance of machine-authored writes, blast radius of a wrong inference, and when a human must be in the loop. The words transaction, ACID, rollback, provenance, and audit do not appear. The nearest the paper comes is one sentence in §8: "integrating neural components introduces temporal consistency challenges, where model updates may lead to temporary discrepancies between the base graph data and learned representations" — which is about embedding staleness, not about a wrong autonomous DELETE. This is the most important gap in the list, and it is squarely within the paper's own scope.
3. **Internal contradiction between the agentic premise and the system requirements.** §7.1, Scenario Feature (3): "Read intensive workload. During the serving stage, most of the graph data and embeddings are queried more frequently rather than updated[59]." The System requirements in the same subsection are derived from that assumption. But continuous learning through active CRUD (§1, System bullet; §1, final paragraph) is precisely a write path that the classical graph-database workload does not have. Either the agentic write traffic is negligible — in which case say so and quantify — or the read-intensive premise needs revisiting for exactly the system this paper proposes. Right now Challenge 6 is a distributed-NGDB challenge, not a distributed *Agentic* NGDB challenge.
4. **Internal contradiction about neuro-symbolic generalisation.** §4.2, "Neuro-symbolic Methods," closes with: "Notably, the neuro-symbolic approach achieves the same level of generalizability in queries as the classic database research." One subsection earlier, §4.1, "More Advanced Query Types," item (ii): "the AVG and COUNT functions in SQL but not in current CQA models." Aggregation, path, and recursive queries are bread and butter for a classic graph database. The §4.2 sentence also carries no citation. Please either scope it ("the same generalisability *over the EFO-1 family*") or drop it.
5. **Challenge 9 (§10) contains no citations at all.** Not one reference marker appears between the section heading and the end of the "Joint Learning Framework" paragraph — this in a section that

defines a system (NGDB-RAG), attributes properties to it, and states a training objective. Graph-based retrieval augmentation is an active literature with published systems and published failure modes; a section proposing NGDB-RAG needs to say what is new relative to it. As it stands, the most speculative section of the paper is also its least supported.

6. **§10 asserts capabilities that §2 and §4.1 list as unsolved.** §10, "Functionality of NGDB-RAG": NGDB-RAG "supports various query types, including temporal, spatial, and numerical reasoning." §2 (Challenge 1) presents numeric literals and event/temporal semantics as open — "Despite these advancements, challenges remain" — and §4.1 lists function support as a research gap with one preliminary attempt [8]. Spatial reasoning is not discussed anywhere else in the paper. Three pages apart, the same capabilities are an open challenge and a delivered feature.
7. **"Significant gains" without a number or a test** (§3, second paragraph): "Reinforcement Learning from Knowledge Graph feedback (RLF-KG) employed proximal policy optimization (PPO) ... achieving significant gains in metrics like Jaccard similarity and Smatch scores across multiple datasets." No effect size, no dataset named at that point, no baseline, no variance, and RLF-KG is introduced without a citation marker on the sentence that names it (the attribution to [15] must be inferred from the preceding sentence). Either give the numbers with their source table, or write "reported improvements" and cite.
8. **The decode complexity in §6 appears to be over the wrong set.** §6, "Test Time Scaling in Inference Stage" defines data complexity against "the database size $|E|$ (number of the edges)," then states that query embedding and sequence models need " $O(|Q|)$ to encode query information and $O(|E|)$ to decode the answer by embedding comparison." Decoding by embedding comparison scores the query embedding against *entity* embeddings, so the natural bound is over the entity set, not the edge set — and those differ by a large constant factor in every KG the paper cites. Relatedly, the paragraph asserts "The complexity of neural symbolic search is well studied" without a citation on that sentence, and gives only upper bounds for specific algorithms; the hardness of the underlying evaluation problem is not mentioned, which leaves the impression that $O(|E|^{1Q|})$ is an artefact of current methods rather than partly intrinsic.
9. **Challenge 5 treats a hoped-for regularity as an existing one.** §1, Learning and Inference: "The scaling laws of neural query execution (Challenge 5) must be explored." §6 then imports the neural scaling law from [52][53] and says "During the development of the NGDB model, scaling is also a major thread." No evidence is offered that any scaling law has in fact been fitted for query

embedding or neuro-symbolic CQA models, and the standard benchmarks in this area are small and fixed, which is a real obstacle to observing one. State it as the hypothesis it is, and specify the experiment: which model family, which axis, which metric, over what range.

10. **Novelty is not defended against the autonomous-database literature.** "Autonomous query construction plus continuous self-tuning" is the self-driving DBMS agenda; the paper's own reference list contains NeurDB [75] ("An AI-powered Autonomous Database") and nsDB [74], but they appear only once, in §7.2, as "preliminary practice cases in relational databases" for query planning. A reader's first question about Agentic NGDB is whether it is more than NGDB [4] with an LLM agent bolted on top, and the paper never answers it head-on. One paragraph — what the graph setting changes, what the neural storage changes, what remains genuinely open — would settle it.
11. **Reference [4] is used as both a proposal and a shipped system.** §1 introduces it as something that "has been proposed"; §7.1 cites it for a scenario feature; §8 cites it for what NGDB "emphasizes"; §9 says "Recent work[4] introduces a neural graph engine that learns query planning and execution strategies through interactions with Neural Graph Storage." If that engine is implemented and evaluated, say where; if it is a vision, the verbs in §9 should match the verbs in §1.
12. **Challenge 10 (§11) is an applications list, not a challenge.** Three bullets — autonomous data management, personalized recommendations, complex event processing — each stating what Agentic NGDB "can" do, none naming an obstacle, a metric, or an open question. The framing sentence, "The challenges lie in creating systems that leverage their advanced functionalities across diverse applications," is a restatement of the paper's thesis. Two honest options: fold this into the conclusion as motivating applications, and either live with nine challenges or promote a real tenth (the autonomous-write correctness problem in point 2 is the obvious candidate).
13. **Challenge 4 (§5) leans heavily on generic database privacy, and its gap claim sits awkwardly with the authors' own work.** §5.1 recapitulates unauthorized access, insider threats, inference attacks, k-anonymity, l-diversity, encryption, access control, and differential privacy — none of it NGDB-specific. §5.2 then says "The exploration of privacy issues in NGDBs remains largely underdeveloped," while citing [42] (Privacy-Preserved Neural Graph Databases, KDD 2024) and [49] (FedCQA) in the same subsection. Both can be true, but the gap needs to be stated against those two papers specifically: what do they leave open? Compress §5.1 to a few lines and spend the space there.
14. **EFO-k is glossed incorrectly at first use.** §3, bullet "More Complex Structured Hypotheses": "hypotheses should accommodate EFOk(existential first-order logic) and cyclic queries." In §4.1 the paper correctly defines EFO-1 as existential first-order queries with *one free variable*, so the k

indexes free variables, not the logic. The gloss as printed makes EFO-k synonymous with existential first-order logic and drops the parameter's meaning. Also missing a space before the parenthesis.

15. **The abstract lists four of the ten challenges as though they were the list.** "We identify ten key challenges in realizing Agentic NGDBs: semantic unit representation, abductive reasoning, scalable query execution, and integration with foundation models like large language models (LLMs)." Four items follow a colon after the word ten. Add "among them" or name all ten compactly. In the same spirit, the closing of the abstract ("paving the way for adaptable and autonomous data management solutions") and of §12 ("These advancements promise to redefine how we manage, query, and reason over interconnected data for the future") are promotional register for claims the paper has not established; the body is more measured than its frame, which is the wrong way round.
16. **Small things, offered for the copy pass, not as objections.** §7.1, System Requirements (3): "Scalability.." — double period. §7.2, Query-Oriented Distributed Storage: "Practices in graph database community[62][63][64][65] and vector database community concludes valuable principles" — "conclude," and the articles are missing. §5.2, Promising Defenses (3): "Private Distribute Training" → "Distributed." §9, first paragraph: "hypothesis or query grounding (with free variables) [82][20], and entity[83] or event[11][84])." — unbalanced parenthesis. §9, Semantic Granularity: "such events, propositions, etc." → "such as." §8, first sentence: "Neural Graph Databases (NGDB) are another way of the data model that derives the properties from the existing graphs" — garbled; please rewrite. §4.1, EFO-1 paragraph: "To handle new graph-theoretic features which cannot be represented in tree-formed queries." is a fragment. §4.1, More Advanced Query Types: "(ii) Support of functions The corresponding research gap" — missing punctuation after the run-in heading. §1: Azure Cosmos DB is named without a citation while Neo4j [1] and TigerGraph [2] get one. §1, Learning and Inference: Russell & Norvig [5] is an odd support for the definition of execution "within the neural space." §2: KBLRN is spelled KBlrn in reference [6]. Reference [15]: "62st Annual Meeting" → "62nd."

Recommendation

Major revisions — the survey core (§4, §6, §7) is a real contribution and worth publishing, but the paper must state the criterion by which the ten were chosen, give each challenge a measurable success condition, replace the indicative claims about NGDB-RAG and Agentic NGDB with claims the citations actually carry, resolve the contradictions in points 3, 4, and 6, and add the one challenge its own premise demands: the correctness and safety of writes that no human authorised.

Declarations

Potential competing interests: No potential competing interests to declare.