

NATURAL-LANGUAGE ANALYTICS IN ENTERPRISE DATA PLATFORMS: EVALUATING LARGE LANGUAGE MODEL INTERFACES FOR BUSINESS INTELLIGENCE

¹*Anil Kumar Kolla, ²Ameya Kokate

¹Department of IT, Honor Health, USA.

²Department of Data Science and Analytics, HonorHealth, USA.

Article Received on 05/09/2026

Article Revised on 25/09/2026

Article Published on 03/10/2026

*Corresponding Author

Anil Kumar Kolla

Department of IT, Honor Health,
USA.

<https://doi.org/10.5281/zenodo.23120385>



How to cite this Article: ¹*Anil Kumar Kolla, ²Ameya Kokate. (2026). Natural-Language Analytics In Enterprise Data Platforms: Evaluating Large Language Model Interfaces For Business Intelligence. World Journal of Engineering Research and Technology, 12(10), 260-270.

This work is licensed under Creative Commons Attribution 4.0 International license.

ABSTRACT

Natural-language interfaces to enterprise data promise to remove the analyst as a bottleneck between a business question and its answer. Large language models have made such interfaces practically deployable, and major data platforms now ship conversational query capabilities over governed enterprise datasets. However, the evidence base for their enterprise use remains thin: published text-to-SQL benchmarks measure execution accuracy against curated academic schemas, whereas enterprise deployment is constrained by ambiguous business vocabulary, undocumented schemas, row- and column-level access control, and the practical consequence that a confidently wrong answer delivered to a decision-maker is worse than no answer at all.

his paper contributes an evaluation framework for natural-language analytics in enterprise settings. We define a taxonomy of enterprise query classes distinguishing lookup, filtered aggregation, comparative, temporal, and metric-definitional queries; we specify a metric suite extending execution accuracy with semantic-equivalence accuracy, abstention rate, harmful-answer rate, and governance conformance; and we describe a semantic-preparation methodology — curated question-answering scope, certified metric definitions, business glossary binding, and worked example pairs — that treats natural-language readiness as a data product responsibility rather than a model capability. We propose an adoption instrument grounded in established technology-acceptance and information-systems-success

models, and we analyse failure modes with particular attention to the calibration gap between user trust and system reliability.

Index Terms—natural language interfaces, text-to-SQL, business intelligence, large language models, semantic layer, data governance, self-service analytics, user adoption.

I. INTRODUCTION

The recurring complaint about enterprise business intelligence is that the questions are simple and the answers are slow. A business stakeholder wants to know how a metric moved last quarter by region; obtaining the answer requires a ticket, an analyst, a query, a review, and a dashboard change. The queue is the product's real latency.

Natural-language interfaces address this directly. Given a governed dataset and a question in plain language, the system generates and executes a query and returns an answer. Advances in large language models have moved this from research demonstration to shipped platform capability, and enterprise deployments are now common enough that the relevant question is no longer feasibility but evaluation: under what conditions does such an interface produce answers a business can act on, and how would an organisation know?

Existing benchmarks provide a partial answer. Spider^[1] established cross-domain text-to-SQL evaluation; BIRD^[2] extended this to larger, dirtier databases with external knowledge requirements. Studies of LLM text-to-SQL capability^{[3]–[5]} report substantial and improving accuracy. But these benchmarks share assumptions that do not hold in enterprise deployment. Their schemas are documented and their column names are meaningful; enterprise schemas frequently are neither. Their questions are unambiguous; enterprise questions rest on contested definitions ("active customer," "net revenue") that differ by department. Their evaluation is execution accuracy against a gold query; enterprise evaluation must account for whether the user could tell the answer was wrong. And they contain no access control, whereas enterprise queries must return results filtered to the asking user's entitlements.

This paper contributes

- 1) **An enterprise query taxonomy** distinguishing classes with materially different difficulty and risk profiles, enabling disaggregated evaluation.
- 2) **A metric suite** extending execution accuracy with abstention, harmful-answer rate, semantic equivalence, and governance conformance.

- 3) **A semantic-preparation methodology** specifying what must be true of a dataset before it is exposed to natural-language querying.
- 4) **An adoption and trust-calibration instrument** grounded in established IS research.^{[6]–[8]}

Contribution type and scope. This is an evaluation-framework contribution. It defines a query taxonomy, a metric suite, a preparation methodology, and a benchmark construction procedure. It does not report results from executing that benchmark. Section V-B specifies how the corpus should be built and Table II gives the pre-specified reporting template, so that the framework can be applied and its outputs compared across deployments.

II. BACKGROUND AND RELATED WORK

A. Text-to-SQL

Semantic parsing of natural language to SQL has a long history; the field's modern empirical era begins with cross-domain datasets requiring generalisation to unseen schemas. Spider^[1] introduced complex, multi-table queries across many databases with a strict train/test schema split. BIRD^[2] increased realism by using larger databases with noisy values and by requiring external knowledge to bridge the gap between question phrasing and stored representation — a condition much closer to enterprise reality.

Large language models substantially changed performance on these tasks. Rajkumar *et al.*^[3] evaluated the zero-shot and few-shot capability of LLMs on text-to-SQL. Subsequent work improved results through decomposition and self-correction^[4] and through systematic study of prompting and example selection.^[5] Surveys^[9] map the design space of schema linking, decomposition, and execution-guided decoding.

B. Semantic Layers and Metric Definition

Dimensional modelling^[10] and the semantic-layer tradition in BI exist precisely because raw schemas do not encode business meaning. A metric such as "net revenue" is a definitional artefact: which adjustments are subtracted, which currency conversion date applies, which cancellations are excluded. Where such definitions are centralised and certified, a natural-language system can bind to them; where they are not, the system must infer them, and inference over contested definitions produces answers that are syntactically valid and semantically wrong. This is the single most important structural difference between benchmark and enterprise conditions, and it is the reason Section IV treats semantic preparation as a precondition rather than an enhancement.

C. Reliability and Hallucination

Language models produce fluent output regardless of grounding.^[11] In a natural-language BI setting the failure is specific: the system returns a number. Numbers carry unwarranted authority, and unlike prose, a wrong number gives the reader no stylistic cue that something is amiss. Retrieval augmentation^[12] can ground generation in schema documentation and prior validated queries, and execution grounding constrains output to queries that run — but neither addresses the case where a query runs successfully and answers a subtly different question than the one asked.

D. Adoption

Whether a capability is used is separable from whether it works. The Technology Acceptance Model^[6], UTAUT^[7], and the DeLone and McLean information systems success model^[8] provide validated constructs — perceived usefulness, perceived ease of use, facilitating conditions, information quality, system quality, net benefits — appropriate for evaluating deployment outcomes. Section VI proposes their application here, with an added trust-calibration construct.

E. Gap

Published evaluation is model-centric and benchmark-bound. We are not aware of an evaluation framework for natural-language analytics that jointly addresses governed access control, contested metric semantics, abstention behaviour, and the asymmetric cost of confidently wrong answers in decision contexts. This paper addresses that gap.

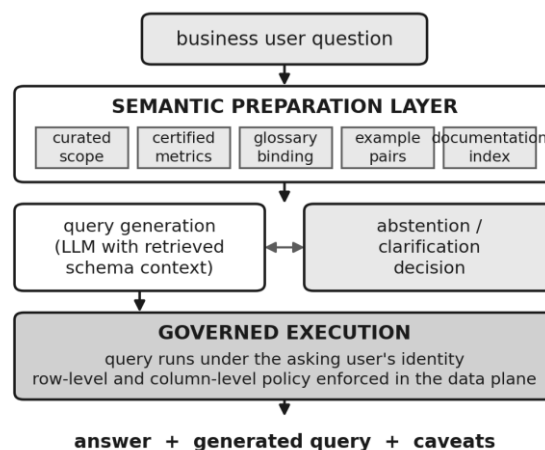


Fig. 1: Reference architecture. Natural-language readiness is located in the semantic preparation layer, and every generated query executes under the asking user's entitlements.

III. Enterprise Query Taxonomy

Aggregate accuracy figures conceal wide variation in difficulty. We propose the following classes for disaggregated reporting.

C1 — Lookup. Retrieval of specific attribute values for identified entities. Low ambiguity; primary failure mode is entity resolution against fuzzy user-supplied names.

C2 — Filtered aggregation. Counts, sums, and averages over filtered populations. The dominant class in practice. Failure modes: incorrect filter semantics, inclusion or exclusion of nulls, and grain errors producing double counting across joins.

C3 — Comparative. Period-over-period, segment-over-segment, or actual-versus-target comparison. Failure modes: misaligned period boundaries, inconsistent denominators, and comparison against a non-comparable baseline.

C4 — Temporal and trend. Queries involving windows, running totals, cohort ageing, or as-of semantics. Highest structural difficulty; failure modes include incorrect window framing and use of recording time where event time was intended.

C5 — Metric-definitional. Queries invoking a named business metric. Difficulty resides not in SQL generation but in definitional binding; these should be answered from certified definitions or refused.

C6 — Out-of-scope. Questions the dataset cannot answer, including causal questions ("why did revenue fall"), predictive questions, and questions about entities outside the governed scope. The correct behaviour is explicit refusal with an explanation, and performance on this class should be reported separately, since a system that never refuses is not safe regardless of its accuracy elsewhere.

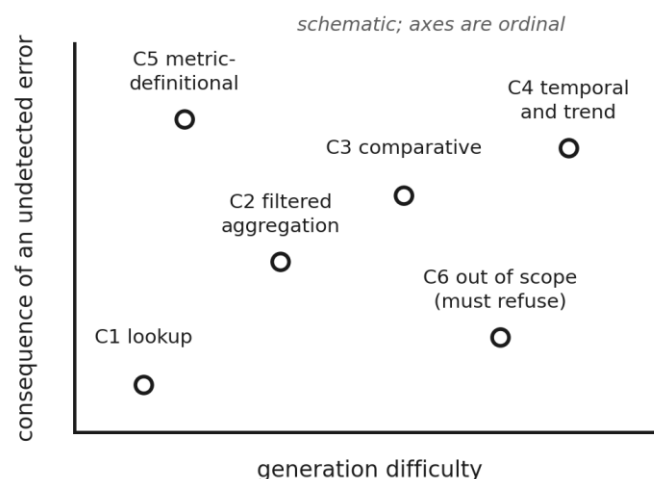


Fig. 2. Schematic positioning of the query classes. Axes are ordinal and indicative; the figure is a design aid, not a measurement.

IV. Semantic Preparation Methodology

The framework's central practical claim is that natural-language readiness is achieved by preparing the data product, not by improving the prompt. We specify five preparation activities.

1. Scope curation. Expose a deliberately narrow, purpose-built set of tables — typically certified gold-layer assets in a medallion-organised lakehouse^[13] — rather than an entire warehouse. Wide schema exposure degrades schema linking and increases the probability of joining tables that should not be joined. The preparation this presupposes is not specific to natural-language systems: metadata-driven ingestion frameworks already centralise column mapping, type conformance, and declarative quality rules as governed configuration rather than per-table code^[14], and the curated, grain-conformant assets they produce are the natural substrate for natural-language exposure.

2. Certified metric definitions. Every metric within scope has a single certified definition expressed as versioned code, with an owner and an effective date. Questions invoking undefined metrics are refused rather than inferred.

3. Glossary binding. A business glossary maps colloquial terms and synonyms to schema elements, including known ambiguous terms flagged for clarification rather than resolution. Where a term is genuinely ambiguous across departments, the system asks rather than guesses.

4. Worked example pairs. A curated set of question–query pairs representative of each query class, used for in-context example selection. Examples should be drawn from real analyst queries with their originating questions, since synthetic examples systematically under-represent enterprise phrasing.

5. Documentation as retrievable context. Column descriptions, grain statements, known data-quality caveats, and coverage limitations are indexed and retrievable.^[12] A grain statement is disproportionately valuable: most double-counting errors originate in an incorrect assumption about table grain.

Governance integration. All generated queries execute under the asking user's identity, inheriting row-level and column-level access policy. The system must not hold elevated privileges of its own. Two derived requirements follow: results must be indistinguishable in form regardless of filtering, so that the presence of restricted data is not inferable from an anomalous total; and refusals due to entitlement must be distinguishable in logs from refusals due to scope, for audit purposes.

V. EVALUATION FRAMEWORK

A. Metrics

Execution accuracy. Proportion of generated queries whose result set matches the gold result set. Reported per query class.

Semantic-equivalence accuracy. Proportion of results that are correct answers to the user's question, assessed by a domain expert. This diverges from execution accuracy in the important case where a query is valid, executes, and returns a defensible but different quantity than intended — the most common enterprise failure and the one benchmarks miss.

Abstention rate and abstention precision. Proportion of questions refused, and proportion of refusals that were correct. A system with high accuracy and zero abstention on a corpus containing C6 questions is misreporting its reliability.

Harmful-answer rate. Proportion of responses that are wrong, unqualified, and plausible — that is, wrong answers a non-expert user would accept. This should be reported as the primary safety metric. It is not the complement of accuracy: a wrong answer flagged with an appropriate caveat, or one so obviously wrong that any reader would reject it, does not count.

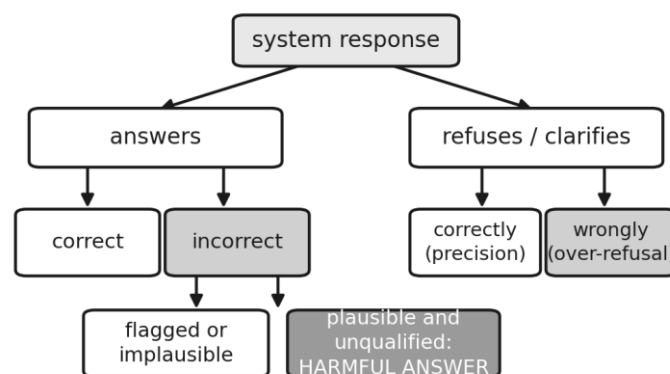


Fig. 3. Response outcome taxonomy. The harmful-answer rate isolates wrong answers that are plausible and unqualified, and is not the complement of accuracy.

Governance conformance. Proportion of queries executing under correct entitlement, and count of any instance in which restricted data was returned or inferable.

Clarification rate. Proportion of ambiguous questions eliciting a clarifying question rather than an assumed interpretation. High clarification rates on genuinely ambiguous input are desirable, not a defect, though they trade against perceived ease of use.

Latency and cost per question.

B. Evaluation Corpus Construction

The corpus should be built from real stakeholder questions harvested from analyst request queues and BI support channels, not authored for the purpose. Each question is annotated with query class, a gold query, a gold answer, and — critically — an ambiguity label indicating whether a single correct interpretation exists. Questions with no single correct interpretation are retained deliberately, since they test clarification behaviour. The corpus must include C6 out-of-scope questions in realistic proportion.

C. Baselines and Ablations

Comparators: (i) analyst-authored queries as the accuracy ceiling; (ii) existing dashboard self-service as the incumbent; (iii) the natural-language system without semantic preparation, isolating the contribution of Section IV; (iv) ablations removing each preparation activity individually.

D. Pre-Specified Reporting Template

Table II. Pre-specified reporting template. Abstention rate and precision, clarification rate, and latency should be reported alongside. Results should be given per query class before aggregation, and dated against both model and schema version.

Configuration	Execution accuracy	Semantic accuracy	Harmful-answer rate	Governance violations
Analyst queries (ceiling)	—	—	—	—
Dashboard self-service	—	—	—	—
LLM with schema only	—	—	—	—
LLM with retrieval	—	—	—	—
Proposed: full preparation	—	—	—	—

VI. Adoption and Trust Calibration

Deployment outcomes should be measured with instruments adapted from TAM^[6], UTAUT^[7], and DeLone and McLean^[8], covering perceived usefulness, perceived ease of use, information quality, and net benefit. To these we add a **trust-calibration** construct measuring the correspondence between user confidence in an answer and that answer's actual correctness.

Trust calibration is the practically decisive variable. A well-calibrated user community extracts value from a moderately accurate system, because members verify answers in the circumstances where verification matters. A poorly calibrated community is endangered by a

highly accurate system, because the residual errors are accepted without scrutiny. Measurement is straightforward in principle: present users with a mixed set of correct and incorrect system answers and record confidence ratings alongside correctness.

Behavioural telemetry complements survey instruments: query volume by user cohort, repeat usage, refinement rate (questions re-asked after an unsatisfactory answer), and escalation rate to human analysts. A falling escalation rate is ambiguous — it may indicate improving system performance or eroding scepticism — and should always be interpreted alongside calibration measurement.

VII. DISCUSSION AND LIMITATIONS

The definitional problem is not a model problem. Where an organisation lacks agreement on what a metric means, no interface can resolve it. Deployment frequently surfaces this disagreement for the first time, which is valuable but is an organisational outcome rather than a technical one.

Expert evaluation limits scale. Semantic-equivalence accuracy and harmful-answer rate require domain expert judgement, which bounds corpus size and introduces inter-rater variability. Inter-rater agreement should be reported.

Non-stationarity. Both the model and the underlying schema evolve. Any reported accuracy is a point measurement; continuous regression evaluation against a held-out corpus is necessary, and results should be dated and versioned against both model and schema.

Generalisation. Results are strongly domain-specific. A system performing well on a curated sales mart tells us little about its performance on clinical or financial reporting data with different structural complexity.

Scope. The framework addresses querying of governed structured data. It does not address analytical reasoning, causal inference, or forecasting, and systems should be constrained to refuse such questions rather than attempt them.

VIII. CONCLUSION

Natural-language analytics is best evaluated not by execution accuracy on benchmark schemas but by whether it produces answers an organisation can safely act upon. This paper proposed a query taxonomy supporting disaggregated reporting, a metric suite foregrounding

abstention and harmful-answer rate alongside accuracy, a semantic-preparation methodology locating readiness in the data product rather than the model, and an adoption instrument centred on trust calibration. The framework's guiding claim is that in decision-support settings the cost of errors is asymmetric, and evaluation designed around average accuracy will systematically overstate readiness.

Future work should develop scalable proxies for semantic-equivalence judgement, study clarification-dialogue strategies that improve correctness without degrading perceived usability, and investigate longitudinal trust calibration in deployed user communities.

REFERENCES

1. T. Yu *et al.*, "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2018; 3911–3921.
2. J. Li *et al.*, "Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs," in *Proc. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.
3. N. Rajkumar, R. Li, and D. Bahdanau, "Evaluating the text-to-SQL capabilities of large language models," 2022; arXiv: 2204.00498.
4. M. Pourreza and D. Rafiei, "DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
5. D. Gao *et al.*, "Text-to-SQL empowered by large language models: A benchmark evaluation," *Proc. VLDB Endowment*, 2024; 17(5): 1132–1145.
6. F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, 1989; 13(3): 319–340.
7. V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, 2003; 27(3): 425–478.
8. W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A ten-year update," *Journal of Management Information Systems*, 2003; 19(4): 9–30.
9. G. Katsogiannis-Meimarakis and G. Koutrika, "A survey on deep learning approaches for text-to-SQL," *The VLDB Journal*, 2023; 32(4): 905–936.

10. R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.
11. Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, 2023; 55(12): 1–38.
12. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
13. M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in *Proc. Conf. Innovative Data Systems Research (CIDR)*, 2021.
14. A. K. Kolla, A. Kokate, R. Vellanki, R. Lolam, and P. Bandi, "A metadata-driven framework for automated bronze-to-silver data ingestion, validation and loading on Databricks," *European Journal of Information Technologies and Computer Science*, 2026; 6(3): 1–7, doi: 10.24018/compute.2026.6.3.70578.
15. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.