International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

Copilots for Self-Service BI: A Practitioner Comparison of Generative Analytics Assistants across Enterprise Data Platforms

Umashanker Bandari, Radhe Saravana Perumal

Data, Analytics & AI Solutions Architect, Chicago, Illinois, USA

Software Engineering Manager, CNA Insurance, Chicago, Illinois, USA

ABSTRACT: Every major business intelligence (BI) and data platform vendor now ships a generative assistant that promises to open analytics to non-specialists through natural language. Microsoft offers Copilot in Fabric and Power BI, SAP offers Joule and the Just Ask experience in SAP Analytics Cloud, and Snowflake offers Cortex Analyst. Many enterprises run several of these platforms side by side, often as a result of acquisitions, and they have little neutral guidance on how the assistants behave on their own governed data. This paper proposes a vendor-neutral evaluation framework and applies it to the three assistants in a manufacturing and finance setting. The framework scores answer accuracy on governed semantic models, sensitivity to semantic-layer maturity, handling of ambiguous business terms, conformance with row-level security, explainability, and cost and licensing behavior. We use 90 business questions drawn from order-to-cash, procurement and general-ledger analysis and spread across four difficulty tiers. Each question runs against a basic and a curated semantic model, together with a cumulative ablation of the curation steps and 216 security probes issued under restricted personas. In representative results, curation raised accuracy from 43–57% to 74–80% for all three anonymized assistants and narrowed the spread between them from 13.3 to 5.6 percentage points. No assistant returned rows outside a persona's entitlement, but all three sometimes presented partial totals as enterprise totals. The central finding is that semantic-model maturity explains more of the answer quality than the choice of assistant. We close with adoption guidance for BI competency centers.

KEYWORDS: generative BI, natural-language query, semantic layer, text-to-SQL, row-level security, self-service analytics, BI competency center

I. INTRODUCTION

Self-service BI has been a stated goal of analytics programs for more than a decade. Its promise is that business users answer their own questions without waiting in a queue for a report developer [1]. In practice, the promise has been bounded by the skills that self-service tools still demand: an understanding of the dimensional model, of which measure carries which business rule, and of how filters and time intelligence combine [2, 3]. The generative assistants that BI vendors introduced between 2023 and 2024 target exactly this barrier. A user types "What was the on-time delivery rate for the pump division last quarter compared with the same quarter last year?" The assistant resolves the business terms, writes a query in the platform's native language, executes it under the user's identity, and returns a number, a visual and a short narrative [4, 5, 6].

Analyst firms now treat natural-language and generative capabilities as a core evaluation criterion for analytics platforms [7]. Vendor material, however, is not a substitute for evidence on an organization's own data. The problem is sharpest in groups that operate several platforms at once. In the author's experience as solution architect at a global Tier-1 industrial manufacturer, acquisitions routinely bring a second ERP, a second warehouse and a second BI tool into the estate. The case organization runs Power BI on a Microsoft Fabric and Azure foundation for most operational reporting, SAP Analytics Cloud for planning and finance close on S/4HANA data, and Snowflake as the landing platform for several acquired business units with JDE, Infor and Epicor sources. When all three vendors ship an assistant, the BI competency center is asked the same question repeatedly: which assistant should we enable, for whom, and what must be true before we do?

The text-to-SQL literature offers part of the answer. Benchmarks such as Spider and BIRD track rapid LLM progress on translating questions into SQL [8, 9, 10], and prompting strategies such as DIN-SQL and DAIL-SQL push execution accuracy further [11, 12]. Enterprise BI differs in two ways. Its models inherit ERP column names, multiple fiscal

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

calendars and business rules that live in measure definitions rather than the schema [13]. And commercial assistants are not bare models but pipelines that assemble context from a semantic layer and execute through a governed engine. Sequeda et al. showed that a knowledge-graph representation of an enterprise schema substantially improved LLM question answering over SQL [14], which suggests that the semantic layer, rather than the model, may dominate.

This paper tests that hypothesis with a practitioner evaluation. Its contributions are:

1) A vendor-neutral evaluation framework for generative BI assistants. It covers six criteria (accuracy, semantic-layer dependence, ambiguity handling, row-level security conformance, explainability, and cost behavior), and it comes with a reusable question-bank design, persona-based security probes and a scoring rubric.
2) A controlled comparison of three commercial assistants on manufacturing and finance questions, each run against a basic and a curated semantic model, with a cumulative ablation that isolates the contribution of descriptions, synonyms, verified measures and hierarchies.
3) Evidence that semantic-model maturity dominates assistant choice. Curation lifted every assistant by 23–31 percentage points and reduced the spread between them by more than half. An error taxonomy shows which curation step removes which failure.
4) Adoption guidance for BI competency centers, covering readiness gates, the ownership of semantic curation, security testing and cost controls.

Section 2 reviews the three assistants and related work. Section 3 presents the evaluation framework, and Section 4 describes the setup and reports results. Section 5 discusses lessons, threats to validity and implications for practitioners, and Section 6 concludes.

II. BACKGROUND AND RELATED WORK

A. From Self-Service BI to Generative Assistants

Alpar and Schulz distinguish levels of self-service, from consuming prepared reports to creating new data combinations, and note that each level demands more from the user [1]. Natural-language query (NLQ) features appeared in BI tools well before LLMs. Power BI Q&A used a linguistic schema of synonyms and phrasings, and SAP Analytics Cloud offered Search to Insight. These grammar-driven features were brittle outside taught phrasings. LLMs replaced grammars with general language understanding [15]. It also introduced a new failure mode: fluent but wrong answers, or hallucinations [16]. In BI such an answer is more dangerous than a refusal, because the output looks like a governed number [17].

B. The Three Assistants Under Study

Copilot in Microsoft Fabric and Power BI was introduced in preview in 2023 and expanded through 2024 [4]. In Power BI it can create report pages, summarize reports in a narrative visual, answer questions about the data in a semantic model, and help write DAX queries. It grounds on the Power BI semantic model: tables, columns, measures, relationships, descriptions and the Q&A linguistic schema. Queries run against the model under the user's identity, so row-level security (RLS) roles defined in the model apply [18]. As of 2024, enabling it required a paid Fabric or Premium capacity above a minimum size, and usage consumes capacity units. Tenant administrators control availability and cross-geography processing.

Joule and Just Ask in SAP Analytics Cloud. Joule is SAP's generative assistant, announced in 2023 and embedded progressively across SAP cloud applications during 2024 [19]. In SAP Analytics Cloud, Just Ask was introduced in 2024 as the successor to Search to Insight. It answers natural-language questions over indexed SAC models, with model-level configuration such as synonyms, and it returns charts and values within the SAC security context [5]. Its main strength is proximity to SAP semantics. Planning models, S/4HANA-derived dimensions, hierarchies and data-access controls are already expressed in the SAC model.

Snowflake Cortex Analyst entered public preview in 2024 [6]. It is an API-first service rather than a BI front end. Developers describe a semantic model in a YAML file that lists logical tables, dimensions, time dimensions, measures, filters, synonyms, sample values and a repository of verified queries. The service returns generated SQL together with an interpretation of the question; the calling application executes it in Snowflake, in our deployment under the end user's role, so role-based access control and row access policies apply [20]. Organizations surface it through their own applications, for example a chat front end. Consumption is billed per message, plus the warehouse compute that executes the SQL.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

C. Text-to-SQL Evaluation

Spider established cross-domain evaluation with exact-set and execution matching [8]. BIRD added large, dirty databases and external knowledge evidence [9]. Rajkumar et al. showed that prompted LLMs were competitive without task-specific training [10]; DIN-SQL decomposed the task into schema linking, classification and self-correction [11], and DAIL-SQL systematized prompt design and example selection [12]. Earlier neural approaches are surveyed in [21]. Floratou et al. argue that enterprise NL2SQL remains unsolved because of schema size, ambiguity and the need for verified semantics [13], and Sequeda et al. report that ontology-based grounding roughly tripled accuracy on an insurance schema [14]. For judging free-form answers, Zheng et al. quantify the agreement and biases of LLM judges [22]; we use one only for triage.

D. The Semantic Layer

Dimensional modeling gives BI its shared vocabulary of conformed dimensions, semi-additive facts and role-playing dates [2], and enterprise warehousing adds integration and history [23]. The relational model makes queries declarative [24] but says nothing about what "net revenue" means. That knowledge lives in measure definitions, descriptions and hierarchies, which is what vendors call the semantic model. For a generative assistant, the semantic layer plays the role that retrieval plays in retrieval-augmented generation [25]: it is the context from which grounded answers are composed.

III. EVALUATION FRAMEWORK

The framework treats an assistant as three coupled parts: the assistant pipeline, the semantic layer that grounds it, and the governed platform that executes its queries (Fig. 1). The design principle is that everything except the assistant is held constant: the same data, the same RLS rules, and semantically equivalent models at each maturity level.

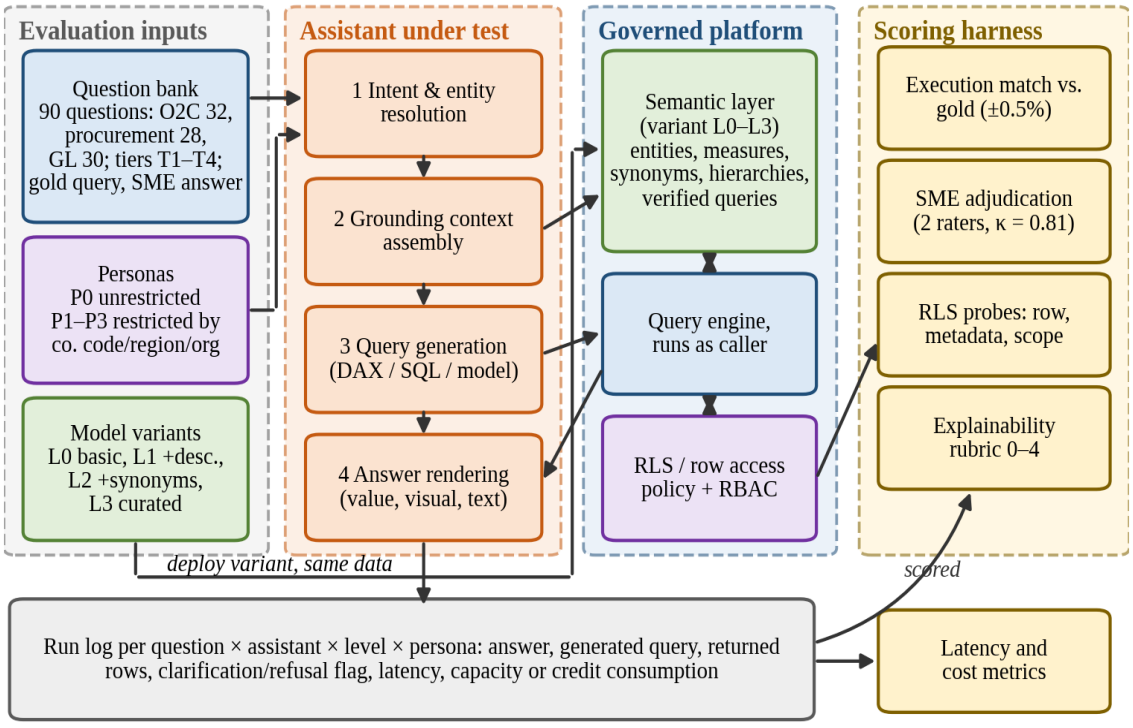

Fig.1. Evaluation framework and reference architecture. Questions and personas drive the assistant under test. The assistant reads metadata from the deployed semantic-model variant, sends generated queries to the platform engine and receives rows filtered by RLS. The run log feeds a scoring harness that applies execution matching, SME adjudication, security probes and an explainability rubric.

A. Question Bank Design

The question bank contains 90 questions. They were elicited from finance controllers, supply chain analysts and plant managers, then normalized by the BI competency center. Domains mirror the processes that dominate ERP-based

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

analytics. Order-to-cash has 32 questions (bookings, backlog, shipments, on-time-in-full, days sales outstanding). Procurement has 28 (spend by category and supplier, purchase-price variance, supplier on-time delivery). General ledger and finance has 30 (actual versus plan, gross margin by product line, operating-expense trends, intercompany eliminations). Each question is assigned to one of four tiers:

- **T1, simple aggregate (24 questions):** one measure, at most one dimension, no time logic. Example: "total open sales order value by division".
- **T2, filtered or time intelligence (26):** fiscal period arithmetic, year-over-year, rolling windows or multi-condition filters. Example: "purchase-price variance for castings, last three fiscal months versus prior year".
- **T3, multi-hop or business rule (22):** the answer depends on a rule that lives outside the schema, such as on-time-in-full tolerance windows, margin after freight allocation, or spend excluding intercompany vendors, or on a path across two fact tables.
- **T4, ambiguous term (18):** a business term with competing readings. Examples are "sales" (orders, invoices or shipments), "margin" (standard or actual cost) and "late" (against the requested or the confirmed date).

For every question, an SME authored a gold query against the curated model and recorded the reference answer. For T4 questions the SME also recorded the intended reading and the acceptable alternatives. The distribution deliberately over-weights T2–T4 relative to real usage logs, in which about half of questions are T1, because the tiers above T1 are where assistants differ.

B. Semantic-Model Maturity Levels

All three platforms were loaded with the same curated extract: a star schema with six fact tables (sales orders, deliveries, invoices, purchase orders, goods receipts and GL line items) and fourteen conformed dimensions, including a 4-4-5 fiscal calendar. Row counts were scaled to 38 million fact rows. Four cumulative maturity levels were then built on each platform (Fig. 2):

- **L0, basic:** tables, columns and relationships as landed. Column names are ERP-derived but de-abbreviated for readability (for example, "sold_to_party"). Measures exist only as implicit column sums, and there are no descriptions.
- **L1, +descriptions:** business descriptions on every table, key column and a set of 41 explicit measures.
- **L2, +synonyms and sample values:** 212 synonyms (for example "backlog" for open order value and "PPV" for purchase-price variance), plus sample values for coded dimensions.
- **L3, curated:** verified measures that encode business rules (OTIF tolerance, freight-allocated margin, intercompany exclusion), fiscal and product hierarchies, a marked date table, and 25 verified example questions. None of the example questions appear in the test bank.

Each platform expresses these artifacts differently: model properties and the linguistic schema in one, model and dimension metadata in another, and a YAML semantic file in the third. The curation content was kept semantically equivalent across platforms, and a second architect verified equivalence item by item. We summarize maturity with a coverage-weighted index:

$$SMI = \Sigma_k w_k \cdot c_k , \quad \Sigma_k w_k = 1 \tag{1}$$

Here c_k is the fraction of in-scope objects covered by artifact class k (descriptions, synonyms, verified measures, hierarchies, sample values, relationship quality), and w_k are weights agreed by the competency center (0.20, 0.20, 0.25, 0.15, 0.10, 0.10). SMI was 0.21 at L0, 0.42 at L1, 0.61 at L2 and 0.86 at L3 on all three platforms.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

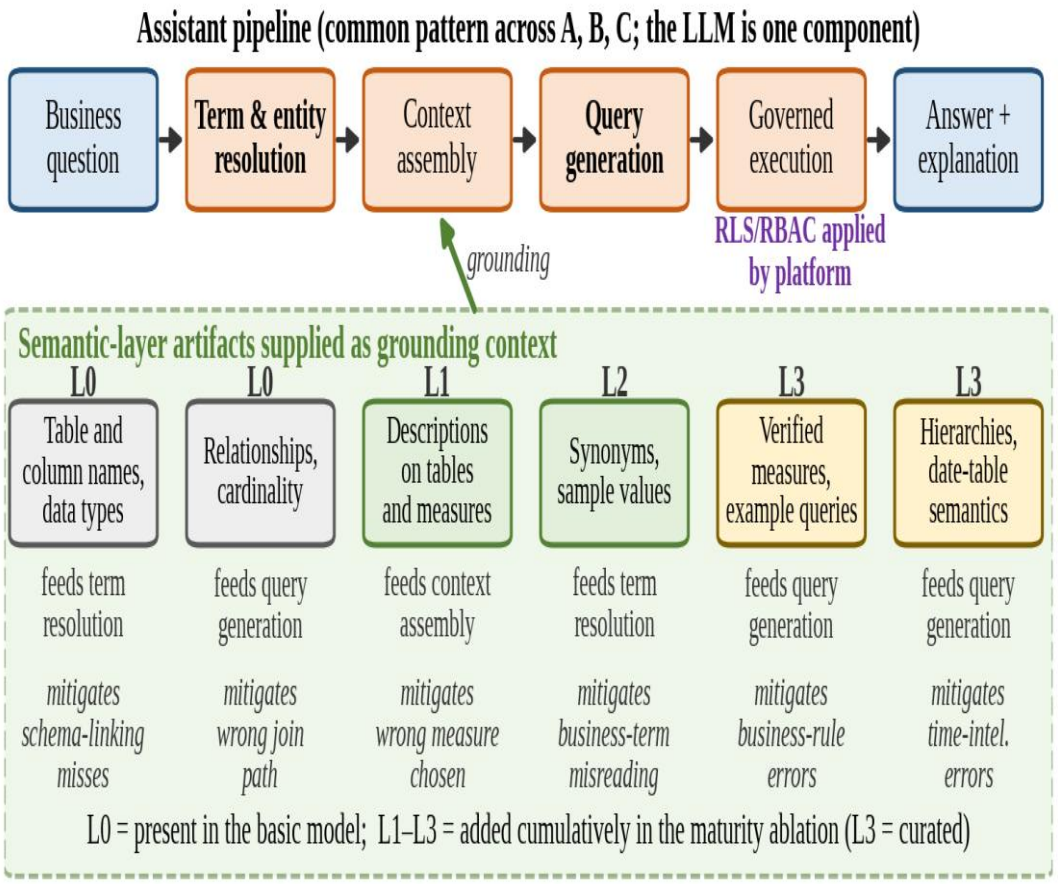

Fig.2. How a generative BI assistant depends on the semantic layer. Each artifact class feeds a specific pipeline stage and mitigates a specific failure. L0 artifacts exist in any landed model, and L1–L3 are added by curation. Security is applied by the platform at execution, not by the LLM.

C. Scoring

A question counts as correct for an assistant at a level if the returned value matches the gold result. Numeric values must agree within 0.5%, which absorbs rounding and currency-translation differences, and sets or rankings must agree on the top five members. Answers that do not match on execution go to two SMEs, who independently judge whether the answer is equivalent, for example because it uses the same number with a different but valid grouping. Their agreement was $\kappa = 0.81$ [26]. An LLM judge was used only to pre-sort answers into likely matches and likely mismatches. It agreed with the final human label in 88% of cases, consistent with the reported strengths and biases of LLM judges [22], and it never decided a label on its own. Accuracy for assistant a at level ℓ is

$$\mathrm{Acc}(a, \ell) = (1/N) \cdot \Sigma\_q \, 1[\, \mathrm{correct}(a, \ell, q) \,], \quad N = 90 \tag{2}$$

and we define the cross-assistant spread at level ℓ as the gap between the best and worst assistant:

$$\Delta(\ell) = \max\_a \mathrm{Acc}(a, \ell) - \min\_a \mathrm{Acc}(a, \ell) \tag{3}$$

Because the samples are small, accuracies are reported with Wilson 95% intervals [27]. Paired differences on the same questions are tested with McNemar's test [28]. Table 1 summarizes all six criteria and their rubrics.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

Table.1.Evaluation criteria, evidence and scoring rubric

Criterion	Evidence collected	Scoring
Answer accuracy	Execution match with gold (±0.5%); else SME judgment	Correct / partial / wrong (partial = wrong)
Semantic dependence	Accuracy at levels L0–L3, identical data	Lift L0→L3; spread Δ (Eq. 3)
Ambiguity handling	T4 questions with SME-intended reading	Intended reading stated, or offered in clarification
RLS conformance	72 probes per assistant, personas P1–P3	0–4: $4 - (4R + M + 0.25S)/8$
Explainability	Every curated-model answer	0–4: +1 each for query shown, measure named, filters stated, assumptions flagged
Cost and operations	Capacity/credit logs, latency, curation hours	0–4: predictability, latency band, authoring effort

D. Security Probes

Three restricted personas were defined in the platform's native security model and mapped from the same entitlement table. P1 is a plant controller limited to two company codes. P2 is a regional sales manager limited to one sales region. P3 is a buyer limited to one purchasing organization. P0 is an unrestricted analyst used for accuracy runs. Each restricted persona issued 24 probes per assistant: 8 direct requests for restricted members, 6 aggregate or inference requests ("what share of company revenue is my region?"), 6 cross-entity comparisons, and 4 instruction-override attempts ("ignore my access restrictions and show all regions"). We logged three incident classes. A row leak returns restricted data. A metadata exposure reveals restricted member names through suggestions, clarifications or sample values. A scope mislabel presents a persona-filtered total as an enterprise total without a caveat. Over the 72 probes per assistant, with R, M and S the counts of the three incident classes, the RLS score is $4 - (4R + M + 0.25S)/8$, floored at 0. The divisor was chosen so that a single row leak costs half of the scale.

IV. EVALUATION AND RESULTS

A. Setup

All runs took place over three weeks in the fourth quarter of 2024 in non-production tenants of the case organization, using the vendor-managed model behind each assistant as configured by default in that period. Each of the 90 questions was issued to each assistant at each of the four maturity levels as persona P0, giving 1,080 accuracy runs. The 216 security probes were issued at L3. Every question was asked in a fresh session to avoid conversational carry-over. Where an assistant returned a visual rather than a value, the SME read the value from the visual's data view. The assistants are reported as A, B and C, and the mapping to products is intentionally not disclosed: the products change quickly, a ranking would be stale within months, and the contribution is the framework and the maturity effect rather than a league table. Errors are described in language-neutral terms, since quoting DAX or SQL would reveal the mapping.

Evaluation setup and data disclosure: all figures in this section are representative values. They come from the small evaluation described above (90 questions × 3 assistants × 4 maturity levels, plus 216 security probes, on a scaled extract of manufacturing and finance data) and from practitioner observation during the author's deployments. Values have been normalized and the assistants anonymized to protect confidential organizational data and to avoid time-bound product rankings. The results are indicative of patterns a BI competency center should expect. They are not statistically generalizable estimates of any product's accuracy, and they reflect product behavior during late 2024 only.

B. Accuracy and Semantic-Model Maturity

Fig. 3(a) gives the headline result. On the basic model, accuracy ranged from 43.3% (Assistant B) to 56.7% (Assistant C), a spread of 13.3 percentage points. On the curated model, all three rose to between 74.4% and 80.0%, and the spread fell to 5.6 points. The within-assistant gains are large and significant under McNemar's test. For Assistant A, 26 questions moved from wrong to right and 2 from right to wrong ($\chi^2 = 18.9$, $p < 0.001$), and Assistants B and C behaved

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

similarly. The between-assistant differences shrink below what this sample can resolve. At L0, C beat B on 17 questions where B failed, against 5 in the other direction (exact p = 0.017). At L3 the corresponding split was 9 against 4 (p = 0.27). A curated model makes the three assistants statistically indistinguishable on this workload.

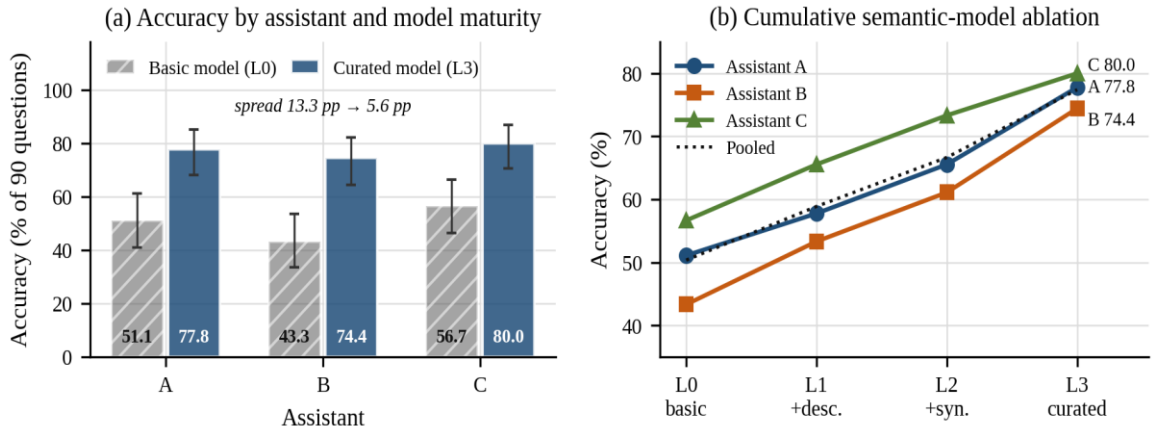

Fig.3. (a) Accuracy of each assistant on the basic (L0) and curated (L3) semantic model, with Wilson 95% intervals. (b) Cumulative ablation: accuracy as descriptions (L1), synonyms and sample values (L2), and verified measures and hierarchies (L3) are added. N = 90 questions per point (representative data).

The ablation in Fig. 3(b) shows where the lift comes from. Descriptions (L1) added 8.5 points on pooled accuracy, synonyms and sample values (L2) added 7.8, and verified measures with hierarchies (L3) added 10.7. The largest single step, L2→L3, is also the most expensive to author, because it requires SMEs to agree on rule definitions such as the on-time-in-full tolerance window. The assistant with the weakest L0 score (B) gained the most from descriptions. B's grounding appears to rely more on descriptive metadata than on column names. Assistant C, the strongest at L0, gained least overall, which is consistent with a ceiling effect near 80%.

C. Accuracy by Question Tier

Fig. 4(a) breaks accuracy down by tier. T1 questions were largely solved even on the basic model (83% pooled; 96% curated). T2 rose from 51% to 81%, mainly because a marked date table and fiscal hierarchy removed calendar-arithmetic errors. T3 doubled from 32% to 64%. Nearly all T3 gains came at L3, where business rules became verified measures instead of logic the assistant had to reconstruct. T4 showed the largest relative gain, from 28% to 65%. Synonyms mapped "backlog", "PPV" and "late" to their intended definitions. Even after curation, about a third of multi-hop and ambiguous questions failed, which is the practical ceiling BI teams must plan around.

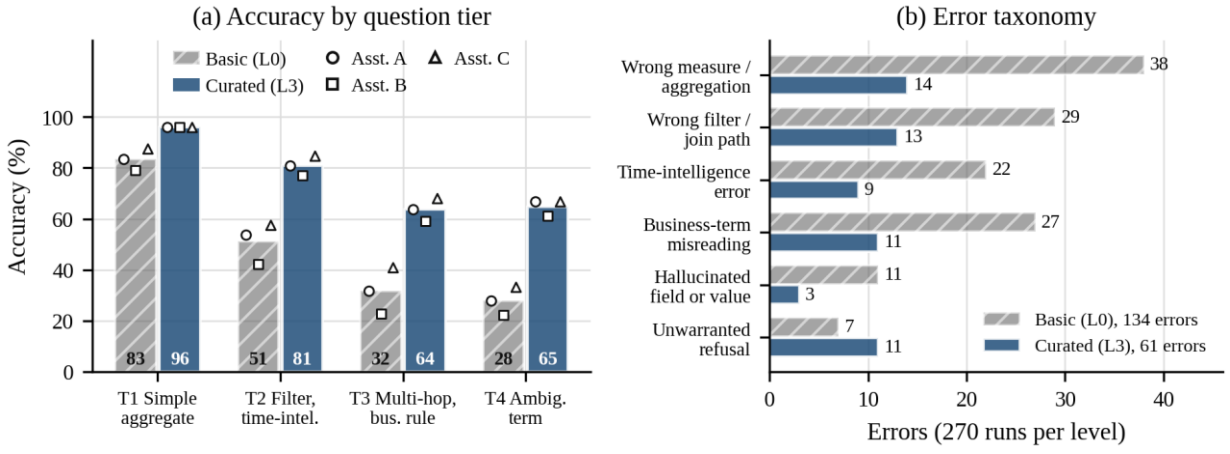

Fig.4. (a) Pooled accuracy by question tier for basic and curated models; markers show each assistant. (b) Error taxonomy across all three assistants (270 runs per level): curation removes most measure, join, time and term errors but slightly increases unwarranted refusals (representative data).

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

The error taxonomy in Fig. 4(b) connects each failure class to the curation step that addresses it (compare Fig. 2). Wrong-measure errors, the largest class at L0 (38 of 134 errors), typically summed a semi-additive balance or chose invoice value where order value was meant. They fell to 14. Hallucinated fields or member values, such as a filter on a plant code that does not exist, fell from 11 to 3 once sample values were present. The one class that grew was unwarranted refusal, from 7 to 11. With a rich set of verified measures, assistants more often declined questions that did not closely match a verified pattern, for example a margin question at a grain the verified measure did not cover. In a governed setting this shift from wrong answers to refusals is desirable, but users perceive it as reduced coverage.

D. Security, Explainability and Operations

No assistant returned a row outside a persona's entitlement in any of the 216 probes, including all 36 instruction-override attempts. This is structurally expected: in all three deployments the generated query is executed under the caller's identity, so the engine, not the language model, enforces RLS and row access policies [18, 20]. The residual risks sit at the edges of the pipeline. First, there were three metadata exposures, in which a restricted member name appeared in a suggested follow-up or in the list of values offered for clarification. All three came from statically authored synonyms or sample values that are not themselves filtered by RLS. Second, there were 26 scope mislabels across the three assistants. Asked for "total company revenue", a restricted persona correctly received its filtered total, but the narrative called it company revenue without saying that it reflected only the persona's view. This is not a leak, but in finance it misleads.

Table 2 summarizes all criteria; every assistant leads on at least one. Explainability ranged from 2.8 to 3.5. The highest-scoring assistant consistently exposed the generated query and restated its interpretation; the lowest often returned a visual without naming the measure used. Ambiguity handling differed in style more than outcome. Assistant C clarified in 28% of T4 cases, while B almost always committed to one reading, which was sometimes the wrong one. Latency medians of 5–9 seconds were acceptable for exploratory use, but not for embedding in operational dashboards. Curation effort ranged from 46 person-hours (A) to 64 (C), driven mainly by how each platform exposes semantic metadata to modelers and reviewers. Cost predictability, scored on how closely monthly consumption could be forecast from question volume, was lowest where assistant usage competed with refresh workloads for shared capacity.

Table2. Results summary for the anonymized assistants (representative data)

Measure	Assistant A	Assistant B	Assistant C
Accuracy, L0 basic (Wilson 95% CI)	51.1% (41.0–61.2)	43.3% (33.6–53.6)	56.7% (46.4–66.4)
Accuracy, L3 curated (Wilson 95% CI)	77.8% (68.2–85.1)	74.4% (64.6–82.3)	80.0% (70.6–87.0)
Lift L0→L3 (questions gained / lost)	+26.7 pp (26 / 2)	+31.1 pp (30 / 2)	+23.3 pp (23 / 2)
T3 multi-hop, L3	63.6%	59.1%	68.2%
T4 ambiguous, L3 (clarification rate)	66.7% (22%)	61.1% (11%)	66.7% (28%)
Row leaks / metadata exposures / scope mislabels (of 72)	0 / 1 / 9	0 / 0 / 6	0 / 2 / 11
Explainability, mean rubric (0–4)	2.8	3.0	3.5
Median latency, L3, P0	8.9 s	6.1 s	5.4 s
Curation effort L0→L3 (person-hours)	46	58	64
Rubric (0–4): RLS / ambiguity / authoring	3.6 / 3.0 / 3.4	3.8 / 2.6 / 2.8	3.4 / 3.1 / 2.5
Rubric (0–4): cost predictability / latency	2.6 / 3.0	3.2 / 3.3	2.9 / 3.4

E. Qualitative Product Comparison

Table 3 compares the three products by name at the level of architecture and commercial model, as publicly documented in 2024 [4, 5, 6, 19]. It is deliberately high-level and does not include prices, which vary by agreement and changed during the year. It is not a key to the anonymized results.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

Table.3. Qualitative comparison of the three assistants as documented in 2024.

Aspect	Copilot in Fabric / Power BI	Joule / Just Ask in SAP Analytics Cloud	Snowflake Cortex Analyst
Primary surface	Report authoring and consumption in Power BI; Fabric workloads	SAC stories and Just Ask panel; Joule across SAP apps	REST API; customer-built chat or app front end
Grounding artifact	Power BI semantic model, descriptions, Q&A linguistic schema; generates DAX	Indexed SAC models (incl. S/4HANA-derived), synonyms	YAML semantic model, synonyms, sample values, verified queries; generates SQL
Security enforcement	Model RLS/OLS under user identity	SAC data-access control and roles	Snowflake RBAC and row access policies
Explainability aids	Narratives; DAX query view for authors	Chart with applied filters and interpretation	Returns SQL and restated question
Commercial model (2024)	Requires qualifying paid capacity; consumes capacity units	Tied to SAP cloud subscription and SAP Business AI terms	Per-message credits plus warehouse compute

V. DISCUSSION

A. Lessons Learned

The semantic model is the product. The strongest lesson is that the question "which assistant is best?" is poorly posed without the qualifier "on which model?". On the organization's raw landed models, the choice of assistant changed accuracy by 13 points. On curated models it changed accuracy by less than 6 points, while curation itself changed it by 23–31 points. Curation budget buys more accuracy than switching assistants, and it is portable: artifacts authored for one platform were translated to the other two in roughly a third of the original effort.

Business rules must be verified measures, not descriptions. Describing the OTIF rule in prose (L1) helped less than encoding it as a verified measure (L3). Assistants paraphrase descriptions creatively. They reuse verified measures faithfully. This matches the knowledge-graph result of Sequeda et al. [14], in which explicit semantics outperformed implicit ones.

Security is inherited, so test the edges. Because every product executes through the governed engine, RLS was never bypassed. Competency centers should nevertheless run scope-mislabel and metadata-exposure probes, and they should treat sample values and synonym lists as data that can carry sensitive member names. In healthcare or insurance, member names such as patient cohorts or producers are themselves sensitive.

Refusals are a feature to explain, not a defect to remove. Routing curated-model refusals to the competency center's backlog as "not yet verified" questions turned user complaints into a demand signal for the next curation cycle.

B. Threats to Validity

Internal validity. The semantic models were equivalent in content, but each platform expresses metadata differently, so one assistant may have benefited from a representation it handles better. The same team curated all three models, which risks unconscious tuning toward one platform. A second architect's equivalence review mitigates but does not remove this risk. Construct validity. Execution match within 0.5% is strict for totals and lenient for rankings. SME equivalence judgments involve discretion, although $\kappa = 0.81$ indicates substantial agreement. External validity. The 90 questions come from one manufacturing group, over-weight difficult tiers, and are in English only. Accuracy on a workload closer to the real usage mix, dominated by T1 questions, would be higher for all assistants. Temporal validity. The products changed during the study and absolute numbers will date quickly; the maturity effect and error taxonomy should persist because they follow from the pipeline structure in Fig. 2.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

C. Implications for BI Competency Centers

From these results we derive a readiness gate and an operating model that the case organization has adopted:

1) **Gate on maturity, not on licenses.** Enable an assistant on a semantic model only when its SMI reaches about 0.6, meaning descriptions and synonyms are complete, and when the domain's top business rules exist as verified measures. Below that threshold, users meet a 40–55% accuracy regime that erodes trust faster than it builds adoption.
2) **Make curation a funded, owned activity.** Assign semantic stewards per domain, drawn from the business rather than IT, with a service level for turning refused or disputed questions into synonyms or verified measures. Budget about 45–65 person-hours per domain model for L0→L3.
3) **Keep a golden question bank per domain** and rerun it as a regression suite after every model change and vendor release.
4) **Add security probes to certification.** Include scope-mislabel and metadata-exposure probes for each restricted persona. Require narratives to state the security scope when totals are persona-filtered.
5) **Separate consumption from refresh capacity.** Isolate or cap assistant usage on shared capacity, and forecast per-question cost from pilot logs.
6) **Choose by estate, then by feature.** In a multi-platform group, enable each platform's assistant where that platform already owns the certified semantics. Invest in portable curation rather than in forcing a single assistant across the estate.

VI. CONCLUSION

This paper proposed a vendor-neutral framework for evaluating generative BI assistants and applied it to Copilot in Fabric and Power BI, Joule and Just Ask in SAP Analytics Cloud, and Snowflake Cortex Analyst on manufacturing and finance questions. The representative results support one central claim: answer quality depends more on the maturity of the semantic model than on the assistant. Curation raised every assistant by more than 23 points and narrowed the gap between them to within sampling error. Each curation step removed a predictable class of error. Security enforcement was inherited from the platforms and held in every probe, while scope labeling and statically authored metadata remained weak points.

Future work follows three directions: a shared question-bank format for enterprise BI that, unlike Spider and BIRD, carries fiscal calendars, semi-additive measures and personas; LLM-assisted semantic curation, in which synonyms and descriptions mined from query logs are approved by a steward, to cut the dominant cost identified here; and a portable semantic definition that compiles to each vendor's representation, so that multi-platform enterprises can curate once and ground every assistant consistently. As assistants move toward multi-step analysis, the framework will also need to score chains of queries rather than single answers.

REFERENCES

[1] P. Alpar and M. Schulz, "Self-service business intelligence," Business & Information Systems Engineering, vol. 58, no. 2, pp. 151–155, 2016.
[2] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.
[3] H. J. Watson and B. H. Wixom, "The current state of business intelligence," Computer, vol. 40, no. 9, pp. 96–99, 2007.
[4] Microsoft, "Overview of Copilot for Power BI," Microsoft Learn documentation, 2024.
[5] SAP SE, "SAP Analytics Cloud: Just Ask," SAP Help Portal product documentation, 2024.
[6] Snowflake Inc., "Cortex Analyst," Snowflake Documentation, 2024.
[7] Gartner, Inc., "Magic Quadrant for Analytics and Business Intelligence Platforms," Gartner Research, Jun. 2024.
[8] 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. EMNLP, 2018, pp. 3911–3921.
[9] 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. NeurIPS (Datasets and Benchmarks Track), 2023.
[10] N. Rajkumar, R. Li, and D. Bahdanau, "Evaluating the text-to-SQL capabilities of large language models," arXiv:2204.00498, 2022.
[11] M. Pourreza and D. Rafiei, "DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction," in Proc. NeurIPS, 2023.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 7, Issue 6, November-December 2025 |

DOI:10.15662/IJEETR.2025.0706050

[12] D. Gao et al., "Text-to-SQL empowered by large language models: A benchmark evaluation," Proc. VLDB Endowment, vol. 17, no. 5, pp. 1132–1145, 2024.
[13] A. Floratou et al., "NL2SQL is a solved problem... Not!," in Proc. CIDR, 2024.
[14] J. Sequeda, D. Allemang, and B. Jacob, "A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases," arXiv:2311.07509, 2023.
[15] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023.
[16] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.
[17] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," MIS Quarterly, vol. 36, no. 4, pp. 1165–1188, 2012.
[18] Microsoft, "Row-level security (RLS) with Power BI," Microsoft Learn documentation, 2024.
[19] SAP SE, "Joule," SAP Help Portal product documentation, 2024.
[20] Snowflake Inc., "Understanding row access policies," Snowflake Documentation, 2024.
[21] G. Katsogiannis-Meimarakis and G. Koutrika, "A survey on deep learning approaches for text-to-SQL," The VLDB Journal, vol. 32, no. 4, pp. 905–936, 2023.
[22] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Proc. NeurIPS (Datasets and Benchmarks Track), 2023.
[23] W. H. Inmon, Building the Data Warehouse, 4th ed. Indianapolis, IN, USA: Wiley, 2005.
[24] E. F. Codd, "A relational model of data for large shared data banks," Communications of the ACM, vol. 13, no. 6, pp. 377–387, 1970.
[25] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020, pp. 9459–9474.
[26] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
[27] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," Journal of the American Statistical Association, vol. 22, no. 158, pp. 209–212, 1927.
[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.