

FRED-SemBench: Evaluating Semantic Reliability in LLM Access to Macroeconomic Data

Wilson Wang¹ Chandler Han¹ Peter Zhang^{1,*}

¹Kairos-AI

*Corresponding author: founders@kairos-ai.io

Abstract

Access to authoritative financial data does not ensure semantically correct answers: a language model may retrieve a valid observation associated with the wrong series, transformation, unit, or temporal interpretation. We introduce FRED-SemBench, a 200-question benchmark candidate over independently selected Federal Reserve Economic Data (FRED) concepts that separates retrieval failures from semantic data-access failures. We also instantiate Iris, an ontology-backed semantic layer whose versioned concept contracts bind user intent to provider series, transformations, units, temporal semantics, provenance, and declared coverage. A controlled comparison of an unaided model, FRED and Web access, and Iris reveals a reliability trade-off: raw retrieval substantially improves numerical answering while suppressing abstention and reducing clarification accuracy on underspecified queries. Within its pre-declared support set, Iris preserves the numerical benefit, restores clarification behavior, and attains the strongest strict accuracy, suggesting that this trade-off is not intrinsic to data access. These pilot findings motivate evaluating macroeconomic data access jointly in terms of semantic correctness, boundary behavior, coverage, and operational cost.

1 Introduction

Consider the question “What was U.S. inflation in May?” A model may select headline CPI (CPIAUCSL) or an unadjusted series (CPIAUCNS), and may report a monthly change, an annual change, or an index level. Each result can come from FRED and still answer a different question (Federal Reserve Bank of St. Louis, 2026). Macroeconomic question answering is therefore not only a retrieval problem: the retrieved value must satisfy the intended concept, transformation, unit, and time semantics.

Financial QA benchmarks test numerical reasoning over reports and tables (Chen et al., 2021;

Zhu et al., 2021; Chen et al., 2022; Islam et al., 2023). DBQR-QA adds database querying before reasoning, while FinRetrieval compares structured tools with Web retrieval (Nararatwong et al., 2024; Kim and Huang, 2026). EconWebArena evaluates agents on live, authoritative economic websites, including FRED (Liu and Quan, 2025). FRED-MD and FRED-QD instead benchmark forecasting databases (McCracken and Ng, 2016, 2021). None of these lines explicitly tests whether a retrieved macroeconomic value satisfies the intended concept, transformation, unit, and observation semantics.

We treat this gap as both an evaluation problem and a system-design problem. Our architectural hypothesis is that reliable financial data access requires an intermediate semantic layer that resolves natural-language intent into an explicit, inspectable contract before retrieving a value. The contract makes series identity, transformation, unit, temporal semantics, and support boundaries first-class rather than leaving them implicit in model reasoning. Figure 1 summarizes the benchmark construction and our coverage-aware evaluation protocol.

We ask two questions: **RQ1**: How much does FRED and Web access improve macroeconomic QA over an unaided language model? **RQ2**: Within declared coverage, does a semantic layer improve reliability beyond raw data access? Our contributions are:

- FRED-SemBench, 200 questions over 50 concepts, with six task types targeting retrieval and semantic data-access failures, backed by frozen provider evidence and deterministic answer construction;
- a semantic-layer design, instantiated in Iris, whose ontology uses versioned concept contracts to bind intended meaning to provider data and expose unsupported requests; and

- a controlled comparison that separates gains from raw retrieval and semantic resolution on Iris’s pre-declared support set.

2 FRED-SemBench

2.1 Concept selection

We selected and froze 50 U.S. macroeconomic concepts before measuring their overlap with Iris. A documented rubric balances their use in FRED-MD/FRED-QD, policy and public relevance, source diversity, temporal coverage, and potential for semantic confusion. Selecting the concept universe first prevents the benchmark from being defined by the system under evaluation.

2.2 Question and answer construction

For every concept, we downloaded one official FRED CSV response on July 29, 2026 and stored its request identity and SHA-256 digest. A deterministic builder derived two facts per concept—a level and a change—for 100 total facts. Gold values therefore refer to these exact byte-frozen responses, not to any later reconstruction of what a date-only “July 29 vintage” contained. We then generated four questions per concept and distributed the 200 cases across the six task types in Table 1. The design tests not only whether a value can be found, but whether a system retrieves the intended value or appropriately declines an underspecified request. Whereas AmbigQA studies multiple plausible interpretations in open-domain QA (Min et al., 2020), our ambiguity cases target economic measures, series, and reporting conventions.

Each record contains its question, expected answer, task type, primary and confounding concepts, series identifiers, transformation, unit, observation period, source, and snapshot provenance. Gold records are excluded from all model-visible files.

2.3 Answer validation

An offline validator recomputed all 100 facts from the frozen responses and verified every snapshot hash and numeric derivation. Automated QA found no exact normalized duplicates and flagged 15 near-duplicate pairs for manual adjudication. FRED-SemBench remains under construction because independent review of question wording, economic usefulness, units, and source attribution is not yet complete, reflecting broader concerns about benchmark validity (Bowman and Dahl, 2021). Its temporal questions use the July 29 snapshot rather than

Task	<i>n</i>	Primary failure tested
Direct retrieval	40	Missing or invented value
Series disambiguation	40	Wrong concept or series
Transform/unit	40	Level/change/unit mismatch
Snapshot		
temporal	40	Wrong period or snapshot
Cross-series	20	Invalid comparison/composition
Ambiguous/unsupported	20	Answering when clarification is required
Total	200	

Table 1: Task allocation in FRED-SemBench.

reconstructed ALFRED vintages. Because data revisions change the information available at a historical date (Croushore and Stark, 2001), we call these snapshot-aware, not historical-vintage, questions. A later FRED real-time-period query can differ from the archived response because of revisions or observations posted after the within-day download.

3 Iris Semantic Layer and Ontology

Iris places a curated ontology between an agent and its data providers. Its central object is a macroeconomic *concept contract*, not a FRED series identifier. A contract specifies geography, indicator family, measure, unit, frequency, transformation, source, release semantics, temporal defaults, confidence, status, and supported operations. This follows the use of ontologies as explicit specifications of shared conceptualizations (Gruber, 1993). The separation between a conceptual vocabulary and provider mappings parallels ontology-based data access (Poggi et al., 2008; Xiao et al., 2018) and virtual knowledge graph systems such as Ontop (Xiao et al., 2020). Iris adopts this design principle but is not a general-purpose OBDA reasoner. Figure 2 summarizes how the contracts are constructed and organized in the FRED setting.

Contract construction reconciles official indicator definitions and release documentation, FRED series metadata, and task-language aliases. Curated entries are schema-validated before entering a versioned registry, so provider availability remains distinct from an explicit claim of semantic support.

At evaluation time, the registry contained 20 supported concepts and two candidate concepts, including headline and core CPI variants, unemployment, payrolls, initial claims, real GDP growth, retail sales, housing starts, Treasury yields, and PPI. Dif-

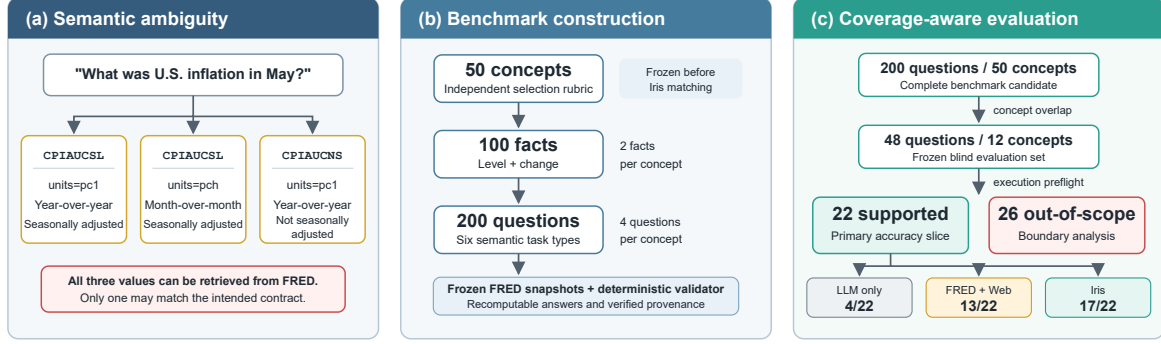


Figure 1: FRED-SemBench construction and evaluation. (a) Requests may denote different series or transformations. (b) The benchmark contains 200 questions over 50 concepts. (c) Twelve concept overlaps yield 48 cases; exact-contract preflight retains 22 and excludes 26 from supported-slice accuracy.

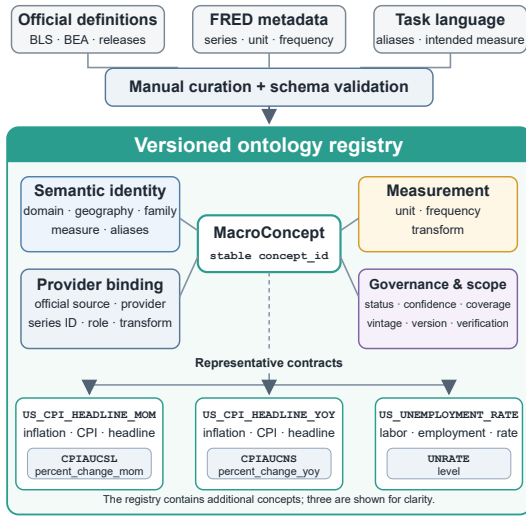


Figure 2: Iris ontology used in the FRED experiment. Schema-validated concept contracts connect semantic identity and measurement to provider bindings, provenance, and coverage; three representative contracts are shown.

ferent contracts can bind the same underlying index to different transformations. For example, month-over-month and year-over-year headline CPI map to transformation-specific provider requests. Iris returns provenance, coverage, confidence, and warnings with the value.

The read-only interface supports concept discovery, contract inspection, value retrieval, semantic resolution, and audit. Provider metadata remains separate from curated meaning: catalog availability does not imply semantic support. Candidate or unsupported requests can therefore be exposed as coverage boundaries rather than coerced into the nearest series.

4 Experimental Setup

Evaluation slices. Twelve of the 50 frozen benchmark concepts overlapped with Iris at the concept level. Their four questions each form a 48-question blind set, but concept overlap does not imply executable support. Before inference, a preflight required exact agreement on concept, provider series, transformation, output unit, and every cross-series input. Twenty-two cases passed (18 numeric and four clarification cases) and constitute the primary accuracy slice; Table 2 shows their concept-level distribution. Thus, 17/22 estimates reliability conditional on exact declared support, not end-to-end macroeconomic QA performance. The other 26 overlap cases are excluded from supported-slice accuracy and characterized separately below.

Conditions. We compare three configurations of the same subscription-backed model (gpt-5.6-sol, medium reasoning, one run per question):

1. **LLM only:** no network, Web search, files, or tools;
2. **FRED+Web:** live Web search plus a read-only wrapper around official FRED GET endpoints; and
3. **Iris:** only the read-only Iris semantic interface, with direct FRED, Web, and shell access disabled.

Each case starts in a fresh thread and temporary runtime without skills, plugins, project instructions, conversation history, or gold fields. The first two conditions completed 48/48 responses. Iris completed 47 directly; one tool-contract violation succeeded on retry and remains visible in the stored

Benchmark concept	In/4	Exclusion flags	Benchmark concept	In/4	Exclusion flags
Headline CPI	2	series, transform	Industrial production	1	series, task, transform
Core CPI	1	cross-input, series, transform	3-month Treasury yield	3	task
Core PCE	2	transform	Credit-card delinquency	3	task
PPI final demand	2	series, transform	Housing starts	1	transform
Unemployment rate	2	cross-input, task, transform	Advance retail sales	2	transform
Initial claims	2	series, transform	Durable-goods orders	1	transform

Table 2: Concept-level exact-contract coverage. Each overlapping concept has four cases; “In” counts preflight-confirmed cases. Exclusion flags can co-occur. Cross-input denotes an uncovered secondary concept in a composed question.

trace. Both provider-backed conditions ran on July 31, 2026. They could request FRED real-time-period semantics through their interfaces, but neither received the benchmark’s archived CSV files.

Scoring. The deterministic scorer requires the expected number at the declared precision, compatible units, and any required clarification. Correct tool use alone receives no credit. Equivalent conversions and paraphrases have not yet been fully adjudicated, so strict scores are conservative. We report no significance test for this single-run pilot. Scoring uses the byte-frozen gold, not the value returned by a live provider at evaluation time. This distinction affected one of the three supported snapshot-temporal cases: both provider-backed responses reported 3.90% for July 28 from a July 29 real-time-period query, whereas the archived July 29 CSV ended at July 27 with 3.96%. The shared mismatch does not explain the four-answer Iris–FRED+Web gap, but absolute scores partly measure reproduction of the archived response as well as semantic selection.

5 Results

RQ1: Data access supplies values and removes caution. FRED+Web increases strict accuracy from 4/22 to 13/22 (Table 3), and the gain is entirely numeric: the unaided model answers 0/18 numeric questions correctly, compared with 11/18 with retrieval. The same change also degrades behavior at the boundary. Safe numeric abstention falls from seven to zero, and clarification accuracy falls from 4/4 to 2/4. The unaided model is not better informed about what it cannot answer, but it does report that it cannot answer; with retrieval it stops. Access therefore converts an unanswered question into a confidently answered one, including when the question was underspecified.

RQ2: The trade-off is not intrinsic. Iris reaches 17/22 strict accuracy, four answers above

Metric	LLM	FRED+Web	Iris
Strict correct	4/22	13/22	17/22
Numeric correct	0/18	11/18	13/18
Clarification correct	4/4	2/4	4/4
Safe numeric abstention	7	0	0

Table 3: Primary results conditional on preflight-confirmed Iris support.

FRED+Web. Numeric accuracy rises from 11/18 to 13/18, and clarification accuracy returns from 2/4 to 4/4. Three quantities move in the same direction, and we read that pattern, rather than the size of any single difference, as the result: with 22 questions, one model and one run, a four-answer gap in strict accuracy is not on its own strong evidence, whereas recovering boundary behavior from 2/4 to 4/4 without regressing numerically is what supports the claim. Conditional on exact declared support, explicit semantic contracts improve both value selection and boundary behavior beyond raw retrieval; this comparison does not establish system-wide superiority.

6 Discussion

Retrieval is necessary but not sufficient. The largest improvement comes from making external data available, but raw retrieval can turn uncertainty into a confidently answered semantic mismatch. Iris’s smaller additional gain matters because it constrains what a retrieved number means. The design implication is not to replace databases or LLMs, but to mediate between them: the ontology defines the intended concept, the provider binding retrieves it, and the coverage contract determines when the system should proceed or expose a boundary. Our evidence supports this approach within Iris’s declared scope; broader claims require wider ontology coverage and additional models.

Coverage and accuracy measure different properties. On all 48 concept-overlap questions, strict

scores are 4/48, 28/48, and 25/48 for LLM-only, FRED+Web, and Iris, respectively. This reversal is expected and informative: 26 cases are outside the frozen Iris execution contract. Raw tools cover more questions, whereas Iris is more accurate on the 22 it claims to support. Reporting only 17/22 would hide limited coverage; reporting only 25/48 would hide the semantic layer’s in-scope behavior. Both quantities are required for a fair comparison.

Excluded cases expose heterogeneous boundaries. Table 4 replaces a single out-of-scope score with the reasons and observed behavior behind it. Transform mismatch dominates, often together with a different provider series; four tasks are inapplicable and two composed questions require an uncovered secondary concept. Iris nevertheless produces eight strict numeric matches, which can arise from valid composition beyond the frozen exact contract, and safely abstains five times. The remaining 13 responses are neither strict matches nor safe abstentions under the deterministic scorer. Because these numeric questions have no pre-registered label declaring refusal correct, this is a boundary characterization, not boundary accuracy: a correct answer does not expand declared coverage, and a refusal is not automatically credited.

Reliability has a cost. The intervention is not free. Across 48 questions, FRED+Web used 149 tool calls and approximately 37 model-minutes, whereas Iris used 230 calls and 130 model-minutes; LLM-only used no tools and 18 model-minutes. Iris’s present CLI composition trades latency and interaction cost for in-scope reliability. Reducing redundant contract discovery is therefore a practical priority.

7 Conclusion

FRED-SemBench evaluates whether a macroeconomic answer has the intended concept, series, transformation, unit, and time contract. Raw FRED and Web access resolves the main retrieval bottleneck; within its declared coverage, Iris adds four strict answers and restores correct clarification behavior. This pilot suggests that reliable financial data access should place an explicit semantic layer and ontology between language models and data providers, while evaluations report semantic correctness, coverage, and cost together.

Preflight flag or observed outcome	<i>n</i>
Transform mismatch	21
Series mismatch	9
Task not applicable	4
Uncovered cross-series input	2
Strict numeric match	8
Safe numeric abstention	5
Other non-matching/pending response	13

Table 4: Characterization of the 26 overlap cases excluded by exact-contract preflight. Flag counts are unique cases and can co-occur; observed outcomes are mutually exclusive and sum to 26.

Limitations

The primary slice contains 22 questions, one model, and one run, so these results are indicative rather than conclusive, and the comparison is confined to the overlap between the frozen concept universe and Iris. Strict scores await independent semantic adjudication, and the full 200-question set still requires domain and licensing review. Because the retrieval conditions did not receive the byte-frozen CSVs, snapshot availability and later revisions cannot be fully separated from semantic error in the absolute scores.

Data and Code Availability

The questions, prompts, outputs, credential-free traces, hashes, and scoring code are publicly available in the versioned GitHub artifact at <https://github.com/wjh19990923/fred-sembench/releases/tag/v1.0.0-paper>. The structured dataset is available on Hugging Face at <https://huggingface.co/datasets/wangjinh/FRED-SemBench>.

References

- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.
- Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [Con-](#)

- vFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292.
- Dean Croushore and Tom Stark. 2001. A real-time data set for macroeconomists. *Journal of Econometrics*, 105(1):111–130.
- Federal Reserve Bank of St. Louis. 2026. Federal reserve economic data (FRED). <https://fred.stlouisfed.org/>. Accessed 2026-07-29.
- Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Eric Y. Kim and Jie Huang. 2026. FinRetrieval: A benchmark for financial data retrieval by AI agents. *arXiv preprint arXiv:2603.04403*.
- Zefang Liu and Yinzhu Quan. 2025. EconWebArena: Benchmarking autonomous agents on economic tasks in realistic web environments. *arXiv preprint arXiv:2506.08136*.
- Michael W. McCracken and Serena Ng. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.
- Michael W. McCracken and Serena Ng. 2021. FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review*, 103(1):1–44.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5783–5797.
- Rungsiman Nararatwong, Chung-Chi Chen, Natthawut Kertkeidkachorn, Hiroya Takamura, and Ryutaro Ichise. 2024. DBQR-QA: A question answering dataset on a hybrid of database querying and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15169–15182.
- Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Linking data to ontologies. In *Journal on Data Semantics X*, volume 4900 of *Lecture Notes in Computer Science*, pages 133–173. Springer.
- Guohui Xiao, Diego Calvanese, Roman Kontchakov, Domenico Lembo, Antonella Poggi, Riccardo Rosati, and Michael Zakharyashev. 2018. Ontology-based data access: A survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5511–5519.
- Guohui Xiao, Davide Lanti, Roman Kontchakov, Sarah Komla-Ebri, Elem Güzel-Kalaycı, Linfang Ding, Julien Corman, Benjamin Cogrel, Diego Calvanese, and Elena Botoeva. 2020. The virtual knowledge graph system Ontop. In *Proceedings of the 19th International Semantic Web Conference*, volume 12507 of *Lecture Notes in Computer Science*, pages 259–277. Springer.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287.