Journal of Data Science 24 (3), 482–503
July 2026

DOI: 10.6339/26-JDS1230
Statistical Data Science

Leveraging Survey Metadata for LLM Reasoning via Knowledge Graphs

Irina Belyaeva[1,*], Christopher Carino[1], and Liang-Chi Wang[1]

[1]*Research and Methodology Directorate, Center for Enterprise Dissemination, U.S. Census Bureau, Suitland, Maryland 20746, United States*

Abstract

Statistical survey metadata contains essential contextual information that underpins the accurate interpretation, discovery, and reuse of statistical data. However, traditional metadata formats are not optimized for consumption by large language models (LLMs), which increasingly function as interfaces for data exploration, question-answering, and decision support. This work introduces a knowledge graph-based approach to modeling survey metadata using semantic web standards and linked data principles, specifically designed to make metadata machine-understandable and LLM-compatible. The core metadata entities, including surveys, datasets, variables, concepts, populations, and provenance, are modeled as rich interlinked nodes that allow reasoning, contextual enrichment, and structured prompting. The graph integrates established ontologies such as the Resource Description Framework (RDF) to promote interoperability and alignment with global standards. We demonstrate how this structure allows LLMs to surface relevant metadata, ground their outputs in authoritative sources, and generate semantically precise responses. This approach enhances transparency, facilitates metadata reuse, and supports the development of artificial intelligence (AI) applications powered by statistical products.

Keywords *large language models; linked data; link prediction; metadata interoperability; retrieval-augmented generation; semantic search; statistical knowledge graphs*

1 Introduction

Recent large language models (LLMs) such as BERT (Devlin, 2018), RoBERTa (Liu et al., 2019), and LLaMA (Grattafiori et al., 2024) demonstrate strong performance across diverse natural language processing (NLP) tasks. Instruction-tuned systems like GPT-4, Claude, and Gemma (Kevian et al., 2024; Team et al., 2024) show substantial potential in complex applications, including education, code generation, and recommendation (Malinka et al., 2023; Li et al., 2022; Liu et al., 2023).

Despite these successes, LLMs are frequently criticized for limited factual grounding. They memorize facts present in the training corpus (Petroni et al., 2019), yet multiple studies show they often fail to reliably recall those facts and may hallucinate—producing statements that are factually incorrect (Ji et al., 2023; Bang et al., 2023). Existing models learn useful linguistic patterns from unlabeled text (Liu et al., 2019), but they struggle with factual knowledge that is sparse, heterogeneous, and embedded in complex forms (Petroni et al., 2019; Logan et al., 2019). As largely black-box models, LLMs also lack interpretability: knowledge is represented

*Corresponding author. Email: irinabelaeva@gmail.com or irina.belyaeva@census.gov.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received September 9, 2025; Accepted April 9, 2026

implicitly in parameters, making it difficult to inspect or validate. The internal patterns and functions used to generate predictions are not readily accessible or explainable to humans (Logan et al., 2019). Even when models produce chain-of-thought rationales (Wang et al., 2023b), those explanations can themselves hallucinate (Golovneva et al., 2023). These limitations hinder deployment in high-stakes domains such as medical diagnosis, legal judgment, and social, economic, and environmental decision-making. A related challenge is domain adaptation: LLMs trained on general corpora may not generalize well to specialized domains or newly emerging knowledge without domain-specific data (Wang et al., 2023a).

In survey statistics—where social and economic indicators inform monitoring and business decisions—correct interpretation and responsible reuse of data depend on rich survey metadata that specify what a measurement represents, how it was collected, and under what assumptions it can be used. When LLMs interact with statistical data without access to such metadata, their outputs can be incomplete, imprecise, or insufficiently grounded in authoritative sources. This limitation undermines the reliability of LLMs as interfaces for data discovery, analysis, and decision support in statistical domains.

Most enterprise metadata repositories are primarily relational and optimized for operational management rather than machine-understandable reasoning. As a result, they often lack explicit graph structure, standardized links across survey programs, topics, variables, and measurements, as well as easily accessible provenance needed to ground LLM outputs. These limitations constrain interoperability with modern LLM workflows for semantic search, variable discovery, and question answering over statistical products.

A growing body of work explores how structured knowledge resources can complement LLMs (Luo et al., 2020; Hu et al., 2022, 2023; Ristoski et al., 2019). Knowledge graphs (KGs) represent facts as triples and provide an explicit, structured representation of knowledge, as exemplified by resources such as Wikidata, YAGO, and NELL (Vrandečić and Krötzsch, 2014; Suchanek et al., 2007; Carlson et al., 2010). Prior studies show that KGs provide accurate explicit knowledge (Ji et al., 2021), support interpretable symbolic reasoning (Zhang et al., 2019), evolve as new facts are added (Mitchell et al., 2018), and can be curated for domain-specific accuracy (Abu-Salih, 2021). Accordingly, integrating KGs with LLMs has attracted increasing attention as a way to enrich pre-training, inference, and interpretability (Petroni et al., 2019; Lin et al., 2019; Dai et al., 2021; Liu et al., 2020, 2021).

Despite this progress, statistical survey metadata remains under-served as a first-class domain in LLM–KG integration. We argue that a standards-aligned survey metadata KG should serve as the primary grounding layer for LLMs operating on statistical products. Such a graph models survey metadata and provenance as interlinked nodes and relations, aligns with established statistical standards including the Generic Statistical Information Model (GSIM) (United Nations Economic Commission for Europe (UNECE), 2025) and Statistical Data and Metadata Exchange (SDMX) (International Organization for Standardization, 2013), and is published using semantic web technologies such as RDF (Cyganiak et al., 2014). When integrated into LLM workflows through retrieval-augmented generation (RAG) and graph-aware ranking and embeddings, this grounding layer reduces hallucinations, improves interpretability and traceability, and enables more precise retrieval and trustworthy AI applications in the statistical domain.

In this work, we present a framework for unifying LLMs with statistical KGs to leverage their complementary strengths and mitigate their limitations. Our main contributions are as follows: (1) we present a standards-aligned Statistical Knowledge Graph (SKG) in RDF—aligned with GSIM and SDMX—with vintage-aware versioning and resolvable identifiers; (2) we introduce an LLM-augmented construction pipeline that generates relation-conditioned canonical node

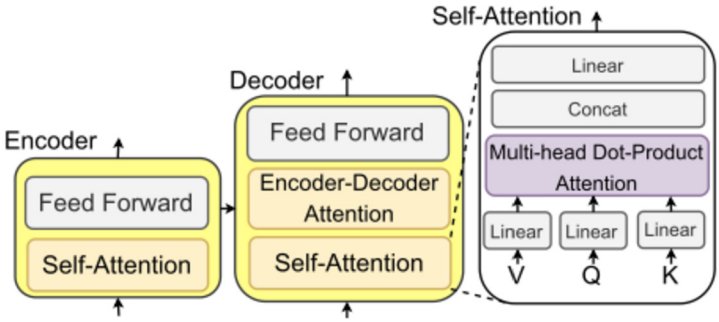


Figure 1: Conceptual illustration of a transformer-based LLM architecture with self-attention, showing encoder and decoder stacks and the multi-head self-attention mechanism.

descriptions and concise statement glosses on salient edges to improve retrieval, disambiguation, and provenance; (3) we propose a fusion architecture that couples SKG-aware retrieval with RAG, including relation-conditioned text encodings that yield graph-aware embeddings and enable grounding-aware prompts with resolvable citations; and (4) we evaluate the approach across analyst tasks—semantic variable retrieval, data-driven topic discovery, and KG link prediction—demonstrating consistent improvements of SKG LLM–RAG over strong text-only and flat-RAG baselines. We conclude by outlining directions for future work in temporal and causal reasoning, external alignment, human-in-the-loop curation, and neural–symbolic inference.

The analyses and views presented in this work are those of the authors and not the U.S Census Bureau.

2 Background

2.1 Large Language Models (LLMs)

LLMs are neural networks trained on vast text corpora to predict the next token in context, and they achieve strong results across many NLP tasks (Yang et al., 2024). This objective yields rich internal representations of syntax, semantics, and world knowledge that transfer to classification, extraction, translation, summarization, question answering, and code generation. As shown in Fig. 1, most modern LLMs are built on the Transformer (Vaswani et al., 2017) architecture, whose self-attention lets the model weigh dependencies among all tokens, enabling parallel training, long-range context handling, and flexible conditioning on prompts, few-shot examples, or retrieved context.

During pretraining, LLMs acquire general language competence from largely unlabeled data. Post-training aligns them with human use: instruction-tuning (Wei et al., 2021; Sanh et al., 2021) teaches them to follow natural prompts; preference optimization (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023) steers outputs toward helpfulness and safety; and system prompting or tool application programming interfaces (APIs) constrain behavior in applications. Because pretrained knowledge is static and incomplete, many systems connect LLMs to external tools—search, databases, code execution, and RAG (Lewis et al., 2020). With RAG, the model conditions on retrieved documents or records, improving factuality and enabling up-to-date answers without retraining.

LLMs excel at language understanding and generation (Brown et al., 2020; Bang et al., 2023), rapid task adaptation via prompting, broad domain coverage, and composing multiple

capabilities (reasoning heuristics, coding, multilingual handling) in a single interface. However, key risks remain: hallucinations (confident but incorrect statements) (Ji et al., 2023), gaps in domain knowledge, sensitivity to prompt phrasing, non-determinism, limited interpretability, and potential bias or privacy concerns. Performance on specialized, structured, or time-sensitive questions often degrades unless outputs are grounded in authoritative sources—motivating our use of a standards-aligned knowledge layer for reliable grounding.

2.2 Knowledge Graphs (KGs)

KGs represent facts as triples (h, r, t) that encode relational facts, where $h, t \in \mathcal{E}$ are the head and tail entities and $r \in \mathcal{R}$ is a relation type drawn from a predefined set. The KG structure supports querying, reasoning, and linking across heterogeneous sources. We group KGs into four broad families, noting typical construction methods, representative examples, and how each can support LLM grounding.

$$\mathcal{KG} = \{(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}, \tag{1}$$

where $\mathcal{E}$ and $\mathcal{R}$ respectively denote the set of entities and relations.

Encyclopedic KGs. General-purpose graphs that integrate broad, real-world knowledge from reference sources and community curation. They offer high coverage, are often multilingual, and are widely reused as background knowledge. For LLMs, they supply common entities and links but may lack the domain precision and provenance required in specialized settings.

Commonsense KGs. Graphs that encode everyday concepts, events, and typical relations—often distilled from text or crowdsourcing. For LLMs, they support implicit assumptions and causal plausibility but do not replace formal, authoritative definitions.

Domain-specific KGs. Focused graphs that capture concepts, constraints, and vocabularies for a particular field (e.g., biomedicine, finance, geology). Examples include the Unified Medical Language System (UMLS) (Bodenreider, 2004) (biomedical), finance (Bennett, 2013), and chemistry ontologies (Hastings et al., 2011). Our work falls here: a statistical survey KG aligned to GSIM/SDMX standards (United Nations Economic Commission for Europe (UNECE), 2025; International Organization for Standardization, 2013) that represents statistical programs, concepts, variables/indicators, and provenance. Compared with encyclopedic graphs, domain KGs are smaller but more precise, auditable, and AI-ready—ideal for grounding LLM outputs in authoritative definitions and lineage.

Multimodal KGs. Graphs that attach or link entities and relations to images, audio, or video; they enable cross-modal retrieval and reasoning (image–text matching, visual question answering, recommendation). For LLMs, they add valuable signals when tasks span text and media.

In summary, encyclopedic and commonsense KGs provide breadth and implicit knowledge, while domain-specific KGs deliver the precision, versioning, and provenance needed for statistical surveys and programs; multimodal KGs extend grounding to settings that involve non-textual modalities. Our work centers on a standards-aligned, domain-specific survey KG published as linked data and fused with LLMs via RAG and graph-aware ranking to produce grounded, auditable answers. Fig. 2 summarizes this integration as a functional abstraction: encyclopedic and domain-specific KGs contribute explicit factual, structural, and domain knowledge, commonsense KGs contribute implicit relational priors and causal plausibility, and multimodal KGs extend the same grounding role to cross-modal settings. KGs ground model outputs with authoritative structure and provenance, while LLMs help populate, refine, and query KGs.

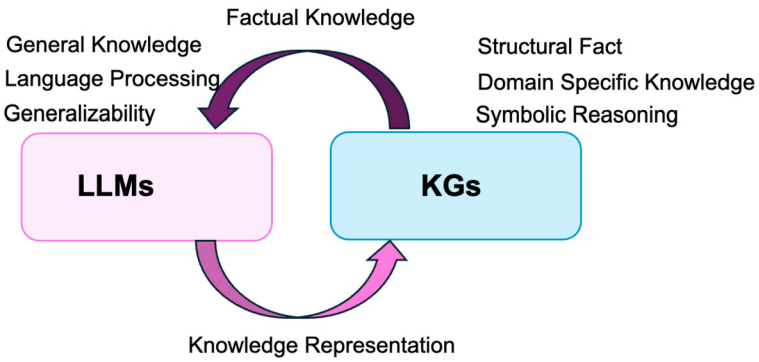


Figure 2: Complementary roles of LLMs and KGs. LLMs provide language understanding and generalization, while KGs supply explicit structural and domain knowledge. The bidirectional loop shows KGs grounding LLM outputs and LLMs supporting KG population, refinement, and querying—together enabling more accurate and interpretable results.

3 Methods

3.1 Data

We use publicly available metadata from the U.S. Census Bureau's Census API (U.S. Census Bureau, 2025). For the American Community Survey (ACS) (U.S. Census Bureau, American Community Survey, 2025), we harvest dataset-level metadata—such as geographic summary levels, groups, and tables—and variable-level metadata, including names, labels, concepts, and data types, via the API discovery endpoints. Because ACS variables evolve across vintages, we retain year-specific metadata and reference both the ACS one-year (U.S. Census Bureau, American Community Survey 1-Year Estimates, 2023) and five-year (U.S. Census Bureau, American Community Survey 5-Year Estimates, 2020) program cycles to support node versioning and stable mappings across releases.

We accessed the API in accordance with the U.S Census Bureau terms of service and citation guidance. The U.S Census Bureau is the source of the original metadata and data.

3.2 LLM-KG Fusion Model Architecture

As shown in Fig. 3, the architecture for LLM and KG data fusion and advanced reasoning consists of four modules: a data layer, an LLM and KG fusion core, a representation learning and RAG module, and an application layer for semantic retrieval, data-driven concept discovery, and KG question answering.

The data layer provides information on programs, variables, and concepts that are harvested and normalized as input to the graph. In the fusion core, LLMs contribute language understanding, generalization, and adaptability, while the KG contributes structured and domain knowledge with clear provenance and interpretability. Two complementary flows operate: facts and provenance retrieved from the graph ground-model output, and model outputs, in turn, help populate, refine, and query the graph. The knowledge extraction and RAG components perform representation learning for entity, relation, and concept extraction, and assemble prompts with retrieved definitions and lineage to support in-context learning and model grounding. The application layer supports semantic information retrieval, data-driven concept discovery, and question answering over the KG.

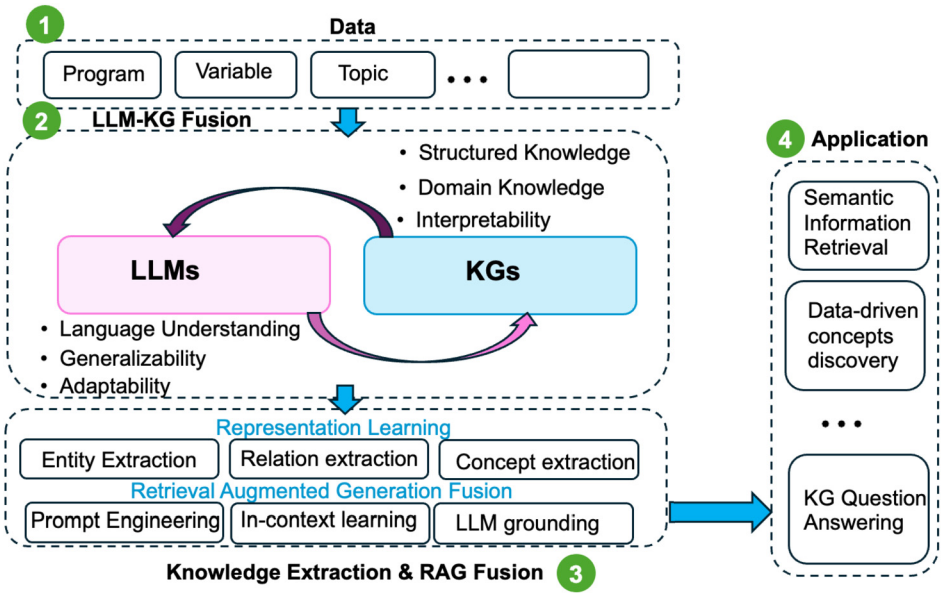


Figure 3: LLM–KG model architecture for advanced reasoning, illustrating the data layer, LLM–KG fusion core, knowledge extraction and RAG integration, and downstream applications.

3.3 Statistical Knowledge Graph (SKG) Model

This section describes the design of the SKG, including its conceptual schema, construction process, node contextualization strategy, and knowledge embedding formulation for downstream retrieval and reasoning tasks. We first describe the conceptual schema of the SKG, followed by its construction, node-level contextualization, and embedding formulation.

3.3.1 Conceptual Model for the SKG

The conceptual model for the SKG shown in Fig. 4 follows a structured and semantically rich organization of metadata entities derived from the American Community Survey via the U.S. Census Bureau's Census API. The SKG is centered around a hierarchical and contextual structure, where entities such as *Program*, *Dataset*, *Variable Group*, and *Variable* are nested, while others like *Concept*, *Universe*, and *Geographical Summary Level* provide semantic annotations and constraints. At the foundation of this model lies the *Program* entity, which represents overarching statistical initiatives such as the ACS. Each program is responsible for producing one or more *Datasets*, which serve as discrete statistical releases corresponding to specific years, formats, or tabulation types. Within each dataset, metadata is further structured into *Variable Groups*, which provide thematic clusters of related variables—for instance, grouping all variables related to educational attainment or housing characteristics. These variable groups serve as organizational containers, each comprising multiple *Variables* that represent individual data points or measurements within the survey. Each variable is semantically enriched through relationships with three additional entities. The *Concept* entity provides interpretive context by defining the meaning or domain of a variable—ensuring that its use in analytical or AI-driven applications is grounded in human-understandable definitions. The *Universe* entity defines the population subset to which a variable applies, such as all households, individuals over a certain age, or employed civilians, thereby setting essential analytic constraints. Additionally, each variable is

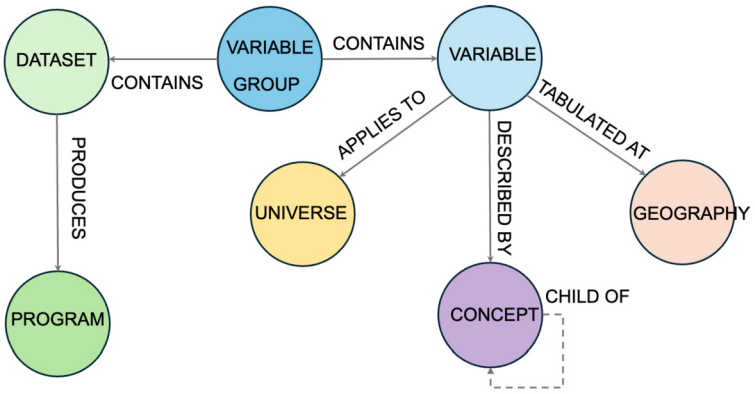


Figure 4: Conceptual model of the SKG showing core metadata entities and their semantic relationships.

linked to one or more *Geographical Summary Levels*, which specify the spatial granularity at which the data is available—ranging from national and state-level aggregates to counties, census tracts, or block groups. These relationships collectively form a graph-based metadata model in which each entity is represented as a node and the connections between them are modeled as typed edges using RDF-based ontologies. This enables inferencing, data discovery, and precise in-context learning by LLMs. For example, when a variable is queried, the graph can surface not just its label, but also its governing concept, its applicable universe, and the geographies across which it is tabulated—all of which can be used to guide reasoning and ensure accurate statistical interpretation. This linked structure transforms traditional metadata into an LLM-compatible format that supports structured prompting, transparent grounding, and machine-understandable lineage—foundational capabilities for AI applications built on statistical data.

3.3.2 LLM-Augmented SKG Construction

We construct a SKG in which each node corresponds to a statistical entity, including programs, variable groups, variables, concepts, topics, and universes. Edges are created from explicit relationships in the metadata or deterministically from the schema. Attributes such as annotations and significance indicators are modeled as separate variable nodes linked by *hasAttribute* property. Provenance is encoded by linking every variable or table node to a program node with properties for vintage year, release type (one year or five year), and version. We represent every fact as a KG triple of the form (h, r, t), where h and t are entities and r is a relation type. In our settings this includes schema facts, crosswalks to external classifications, and year specific instances of variables. Materializing the graph in this way supports both inductive learning, where new variables appear in later releases, and transductive reasoning over relations within a release.

The resulting SKG is rich in provenance and semantic precision and is designed to be ready for AI use. Nodes carry stable identifiers, human readable labels, and machine understandable descriptions, and edges preserve alignment with statistical standards such as GSIM and SDMX. This structure enables retrieval, reasoning, and grounding of LLMs in statistical variable definitions and relationships.

3.3.3 Node Contextualization

We enrich each node in the SKG with a short, canonical description generated by a language model from authoritative metadata. The goal is to provide compact, human-readable context that stabilizes retrieval, improves entity disambiguation across vintages, and supports grounded answers. For every node (program, table or group, variable or indicator, concept, topic, universe) we assemble a deterministic, schema-aware record that includes identifier, label, type, key relations (group, concept, topic, universe, measure), year or vintage, and minimal provenance. The language model receives only this record and, when available, a short definition from source documentation.

3.3.4 Knowledge Embeddings

Given a relational triple (h, r, t), we obtain vector representations by encoding the textual descriptions of the head and tail entities, conditioned on relation r, and by encoding the relation description itself:

$$\begin{aligned} \mathbf{E}_h &= E_s\big(\text{text}_{h,r}\big), \\ \mathbf{E}_t &= E_s\big(\text{text}_{t,r}\big), \\ \mathbf{E}_r &= E_s\big(\text{text}_r\big). \end{aligned} \tag{2}$$

Here, $\text{text}_{h,r}$ and $\text{text}_{t,r}$ are the relation-conditioned descriptions of h and t, and text_r is the description of relation r. The function $E_s(\cdot)$ is a shared text encoder, and $\mathbf{E}_h, \mathbf{E}_t, \mathbf{E}_r \in \mathbb{R}^d$ are the resulting embeddings for h, t, and r.

4 Experiments

In this section we describe the experimental settings for NLP and knowledge extraction tasks. We evaluate the SKG and the LLM–KG fusion on tasks that mirror analyst workflows (see Section 4.1): semantic variable retrieval, KG link prediction, and data-driven concept discovery, with grounding provided by the SKG LLM–RAG model where applicable. We compare the models described in Section 4.3, both with and without the SKG, and report retrieval, link prediction, and topic discovery metrics as defined in Section 4.4. The following subsections describe datasets and query construction, the tasks evaluated and models compared, and the experimental protocol.

4.1 NLP and Knowledge Extraction Tasks

We evaluate the SKG on three tasks that reflect analyst workflows:

1. **Semantic dense retrieval.** Given a natural-language query, return the most relevant statistical variables/indicators together with their metadata.
2. **KG link prediction.** Given a held-out triple (h, r, t), rank the true entity against corrupted candidates to assess whether learned embeddings capture SKG structure.
3. **Data-driven concept discovery.** Given a corpus of variable metadata and a topic prompt template, surface candidate concepts and associated variables that define a coherent topic.

Our hypothesis is that the SKG and its graph embeddings improve retrieval and link prediction quality and enhance data-driven topic discovery relative to unstructured model baselines. We further expect that incorporating RAG on top of the SKG yields more accurate results in settings where it is applied.

Table 1: Models evaluated in the experiments, organized by retrieval strategy, use of structured metadata, and generative capability.

Model	Retrieval Strategy	Structure	Generation
BM25	Lexical	-	-
Text-encoder	Semantic dense	-	-
LLM-prompt-only (parametric)	-	-	✓
LLM–RAG	Semantic dense	-	✓
SKG LLM–RAG	Graph-aware semantic dense retrieval	SKG	✓
BERTopic	-	-	✓

Table 2: Applicability of each model to the evaluation tasks. A check mark indicates that the model is evaluated on the corresponding task.

Model	Semantic Retrieval	Topic Discovery	KG Link Prediction
BM25	✓	–	–
Text-encoder	✓	–	✓
LLM-prompt-only (parametric)	–	✓	–
LLM–RAG	✓	✓	–
SKG LLM–RAG	✓	✓	✓
BERTopic	–	✓	–

4.2 Datasets and Query Sets

We build evaluation sets from the ACS data releases described in Section 2. Each entity in the SKG has a canonical textual description and statement-level gloss as defined in Section 1 and Section 2. We collect natural language queries by paraphrasing table titles, variable labels, and user-facing documentation. For each query, LLM-based annotators mark one or more relevant variables and concepts.

4.3 Experimental Settings

We compare a set of models spanning lexical, dense, generative, and graph-aware approaches to retrieval and topic discovery. The experimental design isolates the contributions of (1) semantic retrieval, (2) LLM-based generation, and (3) grounding in structured statistical metadata via the SKG. Models differ along three orthogonal dimensions: retrieval strategy (none, lexical, dense, or graph-aware), use of structured metadata (absent versus SKG-aware), and generative capability (absent versus LLM-based).

Table 1 and Table 2 summarize the models considered and their applicability to each evaluation task, clarifying how differences in retrieval strategy, structured metadata usage, and generative capability map to the experimental design.

1. **Best Match 25 (BM25) — lexical retrieval**. BM25 (Robertson and Zaragoza, 2009) serves as a lexical baseline over canonical SKG node descriptions. It provides a reference for keyword-based matching without learned semantic representations or generative modeling.
2. **Text-encoder — semantic dense retrieval**. A dense retrieval baseline that encodes

natural-language queries and canonical SKG node descriptions using a shared sentence-level Transformer encoder (Reimers and Gurevych, 2019), ranking entities by vector similarity. This model performs semantic retrieval only and does not use graph structure, relation conditioning, or generation. It isolates the benefit of dense text representations in the absence of explicit graph signals.

3. **LLM–RAG — dense retrieval with unstructured RAG**. An LLM–RAG baseline that retrieves variable metadata from an unstructured corpus and conditions an LLM on the retrieved text. This setting evaluates whether unstructured retrieval improves semantic retrieval and generation without exploiting graph structure or structured statistical provenance.
4. **SKG LLM–RAG — graph-aware retrieval with RAG**. Our full model retrieves SKG entities and statement-level descriptions using graph-aware embeddings and composes grounded LLM–RAG prompts that explicitly encode universe, measure type, vintage, and minimal provenance, with resolvable citations to source nodes or statements.
5. **LLM–prompt–only — parametric generation**. A pure generative baseline in which an LLM produces topic word lists directly from the prompt, relying solely on internal parametric knowledge. The model does not retrieve documents or SKG entities. It does not define a retrieval or ranking function over variables or triples. As a result, it is evaluated only on the topic discovery task and is not applicable to semantic retrieval or link prediction.
6. **BERTopic — topic modeling baseline**. For topic discovery, we additionally include BERTopic (Grootendorst, 2022), a neural topic modeling method that clusters document embeddings to produce data-driven topic representations. BERTopic is evaluated only on the topic discovery task and not on retrieval or link prediction.

The experimental design is structured to isolate the incremental contributions of semantic retrieval, LLM-based generation, and structured statistical grounding. Lexical (BM25) and dense (text-encoder) baselines quantify gains from semantic representations without generation, while generative baselines (LLM–prompt–only and LLM–RAG) assess the effects of parametric knowledge and unstructured retrieval. SKG LLM–RAG integrates graph-aware retrieval with generation to evaluate whether explicit modeling of universe, measure, vintage, and provenance yields additional benefits beyond text-only approaches. To ensure fair comparison, all models use the same query sets and share a fixed prompt budget and, when applicable, the same retrieval depth. Hyperparameters for BM25 and dense retrieval are tuned on a development split, and test-time contexts are restricted to short definitions and minimal provenance so that observed performance differences reflect modeling choices rather than experimental artifacts.

4.4 Evaluation Metrics

We evaluate performance across analyst-relevant tasks using metrics for semantic dense retrieval, knowledge link prediction, and data-driven topic discovery.

4.4.1 Semantic Dense Retrieval Metrics

For semantic dense retrieval, we report Recall@K (3) (Manning et al., 2008) and normalized discounted cumulative gain at K (nDCG@K) (6) Järvelin and Kekäläinen (2002) for variable and concept retrieval, with results reported at $K \in \{1, 5, 10\}$. Recall@K measures whether relevant items are retrieved within the top K results, while nDCG@K captures ranking quality by assigning higher weight to relevant items appearing earlier in the ranked list.

In the metric definitions below, k denotes a generic cutoff; in our experiments, we report

results at fixed cutoffs $K \in \{1, 5, 10\}$.

$$\text{Recall@}k(q) \;=\; \frac{\left|R_q \cap \pi_q^k\right|}{|R_q|}, \tag{3}$$

where R_q is the set of relevant items for query q and $\pi_q^k = \{d_1, \ldots, d_k\}$ denotes the set of the top-k retrieved items.

$$\text{DCG@}k(q) \;=\; \sum_{i=1}^{k} \frac{G\big(r_{q,i}\big)}{\log_2(i+1)}, \tag{4}$$

$$\text{IDCG@}k(q) \;=\; \sum_{i=1}^{k} \frac{G\big(r^*_{q,i}\big)}{\log_2(i+1)}, \tag{5}$$

$$\text{nDCG@}k(q) \;=\; \begin{cases} \dfrac{\text{DCG@}k(q)}{\text{IDCG@}k(q)}, & \text{IDCG@}k(q) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $r_{q,i}$ denotes the relevance grade (binary or graded) of the item at rank i for query q, $r^*_{q,i}$ denotes the relevance grade at rank i in the ideal ranking (items sorted by decreasing relevance), and $G(\cdot)$ is the gain function (we use $G(x) = 2^x - 1$ for graded relevance and $G(x) = x$ for binary relevance).

4.4.2 KG Link Prediction Metrics

For KG link prediction, we evaluate performance using Mean Reciprocal Rank (MRR) (7) and Hits@K (8) on held-out triples under the filtered evaluation protocol. In this setting, when ranking a target entity for a query triple, all other triples known to be true are removed from the candidate set to avoid penalizing the model for predicting alternative correct facts.

MRR captures how highly the correct entity is ranked on average by computing the inverse of its rank, while Hits@K measures the proportion of queries for which the correct entity appears within the top K ranked candidates. We report Hits@K at fixed cutoffs $K \in \{1, 5, 10\}$.

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r_q}, \tag{7}$$

$$\text{Hits@}k = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\big[r_q \leqslant k\big]. \tag{8}$$

Here, Q denotes the set of evaluation queries, r_q is the rank position of the correct entity for query q, and $\mathbf{1}[\cdot]$ is the indicator function. In the definitions above, k denotes a generic cutoff; in our experiments, we report results at fixed cutoffs $K \in \{1, 5, 10\}$.

KG link prediction metrics require models to define an explicit and deterministic scoring function over triples (h, r, t) in order to rank candidate entities.

Text-encoder model. The text-encoder baseline defines a node-only scoring function:

$$f_{\text{node}}(h, r, t) = \text{sim}(\mathbf{E}_h, \mathbf{E}_t)\,, \tag{9}$$

where $\mathbf{E}_e = E_s(\text{text}_e)$ is the embedding of entity e derived from its canonical textual description. This formulation evaluates whether entities that should be linked are embedded close to one another, without explicitly modeling relation semantics.

SKG relation-aware model. The proposed SKG model, defined in (1)–(2), uses a relation-aware scoring function based on relation-conditioned embeddings:

$$f_{\text{SKG}}(h, r, t) = \text{sim}(\mathbf{E}_h + \mathbf{E}_r,\ \mathbf{E}_t)\,, \tag{10}$$

where $\mathbf{E}_h = E_s(\text{text}_{h,r})$, $\mathbf{E}_t = E_s(\text{text}_{t,r})$, and $\mathbf{E}_r = E_s(\text{text}_r)$ are defined as in (2). By conditioning entity representations on the relation, this scoring function explicitly captures the relational structure encoded in the SKG.

BM25 is a lexical text-ranking method and does not explicitly model entities or relations; as a result, it does not define a scoring function over knowledge graph triples and cannot be applied to link prediction. Similarly, LLM–RAG models retrieve and generate text but do not maintain explicit entity or relation representations, nor do they define a deterministic triple-level scoring function. Consequently, neither BM25 nor LLM–RAG is applicable to standard KG link-prediction evaluation, and both are excluded from these metrics.

We therefore report link-prediction metrics only for models that define a deterministic scoring function over triples (h, r, t), namely the text-encoder baseline and the proposed SKG LLM–RAG model.

4.4.3 Data-Driven Topic Discovery Metrics

We assess the quality of discovered concepts and topics using topic coherence measured by normalized pointwise mutual information (NPMI) (12) and topic diversity (TD) (16). Together, these metrics capture semantic consistency within topics and lexical coverage across topics.

Let $W_k = \{w_1, \ldots, w_M\}$ denote the set of the top-M words associated with topic k. Using a reference corpus (e.g., SKG node and statement descriptions together with public documentation), we estimate word and word-pair probabilities by sliding a window of size L over the corpus:

$$p(w) = \frac{C(w)}{N}, \qquad p(w_i, w_j) = \frac{C(w_i, w_j)}{N},$$

where $C(\cdot)$ denotes the number of windows in which a word or word pair occurs, N is the total number of windows, and a small constant ϵ is added in practice to avoid zero probabilities.

Pairwise word association is quantified using pointwise mutual information (PMI) and its normalized variant:

$$\text{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)\, p(w_j)}, \tag{11}$$

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log p(w_i, w_j)}. \tag{12}$$

Topic-level coherence is computed as the average NPMI over all unordered word pairs within a topic:

$$\mathcal{C}_{\text{NPMI}}(k) = \frac{2}{M(M-1)} \sum_{1 \leqslant i < j \leqslant M} \text{NPMI}(w_i, w_j). \tag{13}$$

Corpus-level coherence is obtained by averaging over all topics:

$$\mathcal{C}_{\text{NPMI}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{C}_{\text{NPMI}}(k), \tag{14}$$

where K denotes the total number of topics.

Following prior work, we use $M \in \{10, 20, 25\}$ and window sizes L in the range 10–20 (Bouma, 2009; Newman et al., 2010; Lau et al., 2014; Röder et al., 2015).

To complement coherence, TD quantifies lexical coverage across topics and penalizes repeated top words. Let

$$U = \left| \bigcup_{k=1}^{K} W_k \right| \tag{15}$$

be the number of unique words appearing across the top-M lists of all topics. TD is defined as

$$\mathrm{TD} = \frac{U}{K\,M} \in (0, 1], \tag{16}$$

with larger values indicating fewer repeated words and broader lexical coverage.

For each model, we report corpus-level coherence $\mathcal{C}_{\mathrm{NPMI}}$ (14) and TD (16), together with 95% bootstrap confidence intervals computed over queries or random topic initializations. Higher values indicate better performance for both metrics.

5 Results

In this section, we report empirical results for the evaluation protocols described in Section 4. We evaluate the models introduced in Section 4.3 across the analyst-oriented tasks defined in Section 4.1. Results are organized by task: semantic dense retrieval, KG link prediction, and data-driven topic discovery. The models are evaluated using the metrics defined in Section 4.4. Not all models are applicable to every task. We therefore report results only where a given system defines a meaningful scoring function. The LLM–prompt–only baseline is evaluated only for topic discovery. Unless noted otherwise, scores are macro-averaged over queries to ensure equal weighting across query types, with 95% paired bootstrap confidence intervals. Within each table or figure, the best-performing result is highlighted in bold.

5.1 Semantic Dense Retrieval Task

We evaluate retrieval-capable models (BM25, Text-Encoder, LLM–RAG, and SKG LLM–RAG) on a semantic dense retrieval task in which, given a natural-language query, the objective is to return the most relevant variables together with their associated metadata. Performance is reported using Recall@K and nDCG@K for $K \in \{1, 5, 10\}$. Table 3 reports aggregate results across all cutoffs, and Fig. 5 presents retrieval performance at K=5 and K=10.

Across all cutoffs, the SKG LLM–RAG system consistently achieves the highest Recall@K and nDCG@K among all evaluated models. By retrieving SKG entities and statement-level descriptions that explicitly encode universe, measure, and vintage information, SKG LLM–RAG ranks the intended variable near the top more frequently than LLM-RAG, the dense text-encoder retriever, or BM25. The LLM-RAG baseline ranks second overall, while the text-encoder retriever consistently performs third, outperforming the purely lexical BM25 baseline but falling short of graph-grounded SKG LLM–RAG.

The gains of SKG LLM–RAG are most pronounced for queries that implicitly specify a population or a measure. For example, queries about industry employment for the civilian employed population or the percentage of households with broadband are resolved by matching not only surface terms but also the universe and measure encoded in the SKG descriptions. Queries

Table 3: Semantic dense retrieval results. Metrics are macro averaged over queries; higher is better.

Model	Recall@K			nDCG@K		
	k=1	k=5	k=10	k=1	k=5	k=10
BM25	0.34	0.58	0.68	0.34	0.49	0.54
Text-encoder	0.48	0.73	0.81	0.48	0.65	0.69
LLM-RAG	0.51	0.75	0.83	0.51	0.67	0.71
SKG LLM-RAG	**0.63**	**0.85**	**0.90**	**0.63**	**0.77**	**0.80**

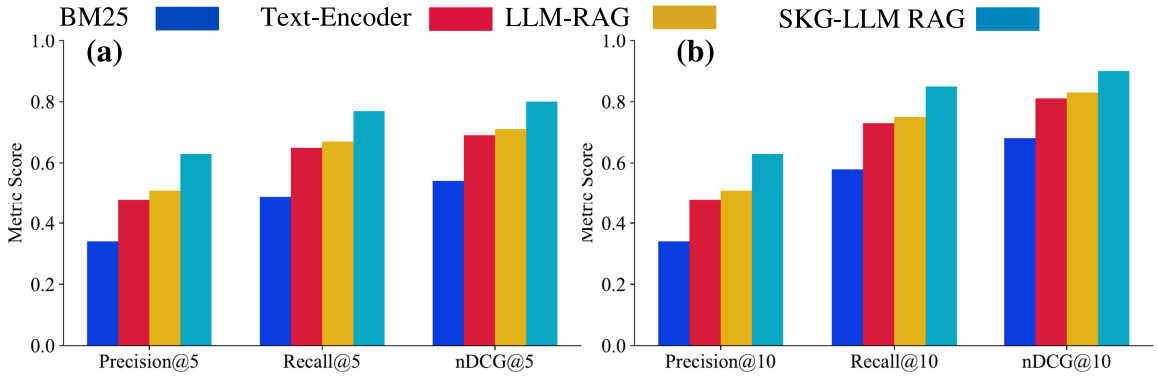


Figure 5: Semantic dense retrieval performance for BM25, Text-Encoder, LLM–RAG, and SKG LLM–RAG across cutoffs K=5 and K=10. Higher is better.

that deviate from official table labels through synonyms or paraphrases similarly benefit from normalized terminology and consistently structured descriptions.

LLM-RAG pipelines can surface passages containing query terms, but they sometimes promote text that does not align with the appropriate universe or vintage. Lexical retrieval such as BM25 performs competitively when queries closely match variable labels, yet degrades when variable names are abbreviated or when multiple vintages share nearly identical titles.

Error analysis identifies three recurrent failure modes in non-SKG systems: short queries that conflate topics and populations, collisions among labels across vintages, and near-duplicate variables that differ only by measure type. Ablation studies confirm that removing SKG node descriptions in favor of bare labels reduces both Recall@K and nDCG@K, particularly for semantically similar variables. Incorporating statement-level glosses on edges—which explicitly restate universe, measure, and year—further improves ranking by providing compact evidence retrievable alongside the entity.

Overall, these results demonstrate that SKG LLM–RAG delivers the most precise and robust semantic retrieval performance for analyst-style queries, particularly in settings where accurate disambiguation requires explicit reasoning over population universe, measure type, and vintage.

5.2 KG Link Prediction Task

We evaluate how well learned representations capture the relational structure of the SKG using KG link-prediction protocols. We evaluate link prediction using two models: a text-encoder

Table 4: KG link prediction on held-out triples. We report MRR and Hits@K for $K \in \{1, 5, 10\}$; higher is better.

Model	MRR	Hits@K		
		k=1	k=5	k=10
BM25	—	—	—	—
Text-encoder (nodes)	0.61	0.46	0.72	0.81
LLM–RAG	—	—	—	—
SKG LLM–RAG	**0.78**	**0.64**	**0.86**	**0.91**

Notes: Link prediction evaluates the ability to recover missing edges using learned KG representations. BM25 and LLM–RAG are retrieval/generation systems and do not define a link-prediction scoring function; we therefore report "—".

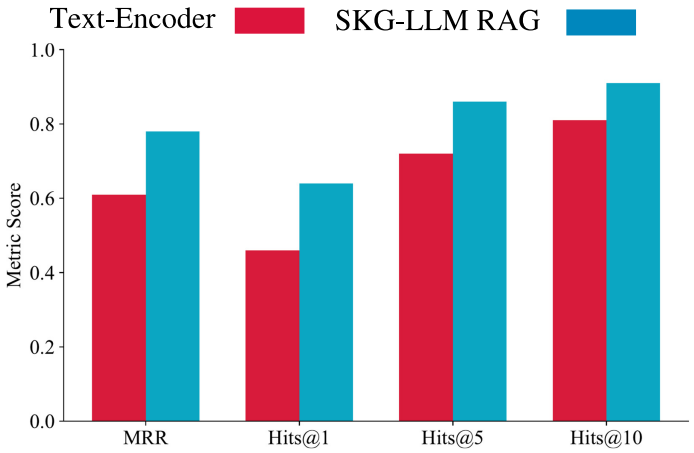


Figure 6: KG link prediction performance. Comparison of link-prediction results for the text-encoder baseline and the SKG LLM–RAG model on held-out SKG triples. Bars report MRR and Hits@K for $K \in \{1, 5, 10\}$ under the filtered evaluation protocol. Higher values indicate better recovery of missing relations.

baseline and the proposed SKG LLM–RAG model. Following common practice, each test triple (h, r, t) is evaluated in a filtered setting. The gold entity is ranked against corrupted candidates generated by replacing h or t with all entities of the same type. Other known true triples are removed from the candidate set. We report MRR and Hits@K for $K \in \{1, 5, 10\}$, as defined in Section 4.4.2.

The text-encoder baseline defines a node-only scoring function (9) based on similarity between entity embeddings derived from canonical node descriptions. In contrast, the proposed SKG LLM–RAG model uses the relation-conditioned descriptions introduced in Section 3.3.4. It defines a relation-aware scoring function (10) that scores candidate triples via similarity between relation-aware entity embeddings and the corresponding relation embedding. Retrieval-based models such as BM25 and LLM–RAG do not define deterministic triple-level scoring functions. Therefore, they are not applicable to link-prediction evaluation. Entries for these models are reported as — in Table 4.

Table 4 and Fig. 6 show that relation-aware SKG embeddings substantially outperform

Table 5: Topic quality for data-driven discovery. We report NPMI topic coherence and TD over the top M=20 words per topic; higher is better.

Model	$\mathcal{C}_{\text{NPMI}}$ (M=20, ↑)	TD (M=20, ↑)
BERTopic	0.200	0.84
LLM prompt-only	0.130	0.72
LLM–RAG	0.220	0.80
SKG LLM–RAG	**0.330**	**0.93**

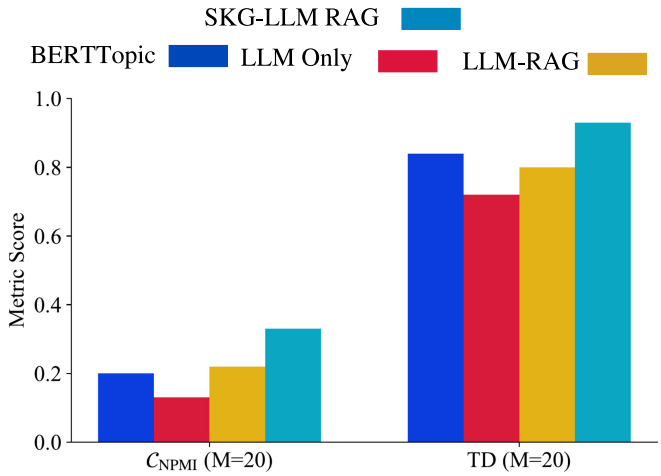


Figure 7: Data-driven topic discovery results. Bars show corpus-level coherence $\mathcal{C}_{\text{NPMI}}$ and Topic Diversity (TD), both computed over the top M=20 words per topic, for BERTopic, LLM prompt-only, LLM–RAG, and SKG LLM–RAG. Higher values indicate better performance.

node-only encodings. The SKG model achieves an MRR of 0.78, compared with 0.61 for the text-encoder baseline. Gains are consistent across Hits@1, Hits@5, and Hits@10. This indicates more accurate ranking of the correct neighbor among hard negatives. Most residual errors arise from aliases and closely related variables that differ only by vintage or minor definitional changes.

Overall, these results demonstrate that relation-conditioned representations enable the SKG LLM–RAG model to capture the SKG's relational structure more effectively than node-only embeddings. Therefore, the SKG LLM–RAG model provides a stronger foundation for downstream retrieval tasks that depend on accurate structural inference.

5.3 Data-Driven Topic Discovery Task

We evaluate systems on a data-driven topic discovery task whose goal is to surface coherent and non-redundant topic candidates from an SKG-aware corpus. We compare four models: (1) BERTopic (Grootendorst, 2022), a strong topic-modeling baseline; (2) an LLM prompt-only baseline without retrieval; (3) LLM–RAG, which retrieves unstructured documentation chunks without using graph structure; and (4) SKG LLM–RAG (ours), which retrieves SKG entities and statement descriptions with explicit universe, measure, and vintage for grounding and provenance. We report topic coherence using corpus-level $\mathcal{C}_{\text{NPMI}}$ (14) and topic diversity using TD (16). Both metrics are computed over the top M=20 words per topic. Table 5 and Fig. 7 present the results.

SKG LLM-RAG achieves the best performance on both coherence and diversity metrics, outperforming LLM–RAG and BERTopic. Gains are largest for prompts that implicitly constrain population or measurement semantics. Prompts about digital access or labor participation yield topics that consistently reflect the correct universe, such as households versus housing units or the civilian employed population, and the intended measure type, such as rates versus counts. Because SKG node and statement descriptions normalize terminology and encode universe, measure, and vintage in a consistent structure, the resulting topics align more closely with underlying statistical concepts and avoid noise from near-duplicate variables.

LLM–RAG improves over the LLM prompt-only baseline but often aggregates passages that mention relevant terms without enforcing scope. Topics may therefore mix universes or vintages, reducing coherence and increasing redundancy. Error analysis reveals two recurring failure modes in non-SKG systems: (1) prompts that conflate related populations, such as households and families, lead to mixed-scope topics; and (2) near-identical labels across releases inflate word repetition and reduce topic diversity. Ablation studies confirm these effects. Removing relation-conditioned node descriptions in favor of bare labels reduces both $\mathcal{C}_{\text{NPMI}}$ and TD. Adding concise statement-level glosses that restate universe, measure, and year improves both metrics by providing compact, retrievable evidence that guides topic grouping.

Overall, these results show that grounding topic discovery in the SKG LLM–RAG pipeline yields topics that are both more semantically coherent and more diverse than those produced by LLM prompt-only or LLM–RAG approaches. Improvements in $\mathcal{C}_{\text{NPMI}}$ and TD indicate that graph-aware grounding provides a stronger inductive bias for assembling interpretable, non-redundant concept sets suitable for downstream retrieval.

6 Discussion

Across tasks that reflect analyst workflows, the SKG improves both retrieval and generation quality when used to ground LLMs. In semantic retrieval, indexing node- and statement-level descriptions yields higher Recall@K and nDCG@K than text-only systems, particularly for queries that implicitly encode population scope or measurement semantics. In data-driven topic discovery, SKG LLM–RAG produces topics that are both more coherent (higher NPMI) and less redundant (higher TD) than BERTopic, LLM–prompt–only, and unstructured LLM–RAG models (Table 5, Figure 7). KG link prediction results on held-out triples further indicate that SKG-aware embeddings capture relational structure, supporting completion of missing edges and multi-hop reasoning. Together, these findings support the central premise of this work: a standards-aligned SKG provides an effective inductive bias for grounding LLMs in statistical settings.

Two design choices appear especially important. First, relation-conditioned node descriptions expose universe, measure type, vintage, and minimal provenance in a compact and consistent schema, stabilizing retrieval and reducing ambiguity when surface forms are similar across years or programs. Second, statement-level glosses attached to edges restate the semantics of key links, enabling retrievers to surface concise evidence alongside entities and allowing RAG prompts to include resolvable resource identifiers. Together, these mechanisms reduce hallucination pressure and improve attribution in downstream question answering.

Grounded retrieval and generation support repeatable analyst workflows, including identifying the correct variable among near-duplicates, tracing the applicable population universe, verifying measure types prior to analysis, and enforcing vintage-specific constraints. The SKG also provides a maintainable path for metadata evolution: new releases are represented as ex-

plicit nodes with versioned lineage rather than silent changes in documentation. This design directly supports auditability and reproducibility.

Finally, aligning the SKG with GSIM and SDMX and publishing it using RDF improves interoperability across statistical programs. It also clarifies governance by treating definitions, universes, and vintages as first-class graph objects with explicit ownership and change histories. This structure supports transparent versioning, deprecation, and crosswalks to external classifications, which are essential for longitudinal and comparative analysis.

7 Conclusion

We presented an industry standards-aligned SKG as a grounding layer for LLMs oriented to consume statistical data. We introduce an SKG that models programs, datasets, variable groups, variables, concepts, universes, and provenance. We enrich nodes with relation-conditioned descriptions and attach statement-level glosses to salient edges. Integrated into RAG, this structure improves semantic retrieval, strengthens topic discovery, and supports link prediction and question answering with resolvable citations.

The results show that SKG LLM–RAG offers consistent gains over text-only and unstructured LLM-RAG pipelines, especially when disambiguation requires an explicit universe, measure type, and vintage. Beyond metric improvements, the approach advances transparency and auditability by grounding answers in authoritative, versioned definitions aligned with GSIM and SDMX.

Limitations and feasibility. Despite these advantages, the proposed approach has several limitations that warrant discussion. Constructing and maintaining a standards-aligned SKG requires upfront investment in metadata engineering, ontology alignment, and governance, which may limit immediate adoption in organizations without mature metadata infrastructure or domain expertise. While the approach is well suited to large statistical agencies and institutions with stable metadata standards, it may be less feasible for ad hoc datasets or domains where definitions, populations, and measures are weakly specified or rapidly changing. Moreover, the benefits of SKG grounding depend on the quality and consistency of the underlying metadata; incomplete or inconsistently versioned metadata can propagate uncertainty into retrieval and generation. Finally, although the SKG improves topic discovery and retrieval, it does not eliminate the need for human oversight, as analysts must still assess the appropriateness of retrieved variables and generated explanations for a given analytical context.

We see five promising directions: (1) temporal and causal reasoning—extending RAG to reason explicitly over time, including definitional changes and population shifts across vintages; (2) entity alignment and external links—developing robust crosswalks to external ontologies and code lists to broaden grounding beyond a single program; (3) human-in-the-loop curation—incorporating active-learning loops in which analysts validate citations and the SKG learns from corrections; (4) neural–symbolic inference—combining text encoders with graph-based retrieval and rule-based constraints for program-specific checks; and (5) evaluation at policy grade—conducting domain-expert assessments of groundedness and utility in realistic decision-support scenarios, beyond standard information retrieval metrics.

Looking ahead, we envision statistical agencies and research teams building and sharing SKGs as reusable infrastructure for discovery, analysis, and communication. With careful governance, privacy safeguards, and human oversight, such graphs can make LLM-powered tools more reliable and interpretable, accelerating responsible use of statistical products in research and decision making.

Supplementary Material

Appendices A-C.

References

Abu-Salih B (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, 185: 103076. https://doi.org/10.1016/j.jnca.2021.103076

Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, ..., Fung P (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (JC Park, Y Arase, B Hu, W Lu, D Wijaya, A Purwarianti, AA Krisnadhi, eds.), 675–718. Association for Computational Linguistics, Nusa Dua, Bali.

Bennett M (2013). The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3): 255–268. https://doi.org/10.1057/jbr.2013.13

Bodenreider O (2004). The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*. 32(suppl_1): D267–D270.

Bouma G (2009). Normalized (pointwise) mutual information in collocation extraction. In: *Proceedings of the Biennial GSCL Conference: From Form to Meaning—Processing Texts Automatically* (C Chiarcos, RE de Castilho, M Stede, eds.), 31–40.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, ..., Amodei D (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka E, Mitchell T (2010). Toward an architecture for never-ending language learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (M Fox, D Poole, eds.), volume 24, 1306–1313.

Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Cyganiak R, Wood D, Lanthaler M (2014). RDF 1.1 concepts and abstract syntax. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. W3C Recommendation. 25 February 2014.

Dai D, Dong L, Hao Y, Sui Z, Chang B, Wei F (2021). Knowledge neurons in pretrained transformers. arXiv preprint.

Devlin J (2018). Bert: Pre-training of deep bidirectional transformers for language understanding/arxiv preprint. arXiv preprint: arXiv:1810.04805

Golovneva O, Chen M, Poff S, Corredor M, Zettlemoyer L, ..., Celikyilmaz A (2023). ROSCOE: A suite of metrics for scoring step-by-step reasoning. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, ..., Ma Z (2024). The llama 3 herd of models. arXiv preprint: arXiv:2407.21783

Grootendorst M (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint: arXiv:2203.05794

Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M (2011). The chemical information ontology: Provenance and disambiguation for chemical data on the biological semantic web. *PLoS ONE*, 6(10): e25513. https://doi.org/10.1371/journal.pone.0025513

Hu N, Wu Y, Qi G, Min D, Chen J, ..., Ali Z (2023). An empirical study of pre-trained language models in simple knowledge graph question answering. *World Wide Web*, 26(5): 2855–2886. https://doi.org/10.1007/s11280-023-01166-y

Hu Z, Xu Y, Yu W, Wang S, Yang Z, ..., Sun Y (2022). Empowering language models with knowledge graph reasoning for question answering. arXiv preprint: arXiv:2211.08380

International Organization for Standardization (2013). Statistical data and metadata exchange (SDMX).

Järvelin K, Kekäläinen J (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4): 422–446. https://doi.org/10.1145/582415.582418

Ji S, Pan S, Cambria E, Marttinen P, Yu PS (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Ji Z, Lee N, Frieske R, Yu T, Su D, ..., Fung P (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38. https://doi.org/10.1145/3571730

Kevian D, Syed U, Guo X, Havens A, Dullerud G, ..., Hu B (2024). Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra. arXiv preprint: arXiv:2404.03647

Lau JH, Newman D, Baldwin T (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539. Association for Computational Linguistics, Gothenburg, Sweden.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, ..., Kiela D (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li Z, Wang C, Liu Z, Wang H, Wang S, Gao C (2022). Cctest: Testing and repairing code completion systems. *2023 ieee/acm 45th international conference on software engineering (icse) (2022)*, 1238–1250.

Lin BY, Chen X, Chen J, Ren X (2019). KagNet: Knowledge-aware graph networks for commonsense reasoning. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K Inui, J Jiang, V Ng, X Wan, eds.), 2829–2839. Association for Computational Linguistics, Hong Kong, China.

Liu J, Liu C, Zhou P, Lv R, Zhou K, Zhang Y (2023). Is chatgpt a good recommender? a preliminary study. arXiv preprint: arXiv:2304.10149

Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA (2019). Linguistic knowledge and transferability of contextual representations. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (J Burstein, C Doran, T Solorio, eds.), volume 1 of *Long and Short Papers*, 1073–1094. Association for Computational Linguistics, Minneapolis, Minnesota.

Liu W, Zhou P, Zhao Z, Wang Z, Ju Q, ..., Wang P (2020). K-bert: Enabling language representation with knowledge graph. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2901–2908.

Liu Y, Ott M, Goyal N, Du J, Joshi M, ..., Stoyanov V (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint: arXiv:1907.11692

Liu Y, Wan Y, He L, Peng H, Yu PS (2021). KG-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In: *Proceedings of the AAAI Conference on Artificial In-*

telligence, volume 35, 6418–6425.

Logan R, Liu NF, Peters ME, Gardner M, Singh S (2019). Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A Korhonen, D Traum, L Màrquez, eds.), 5962–5971. Association for Computational Linguistics, Florence, Italy.

Luo D, Su J, Yu S (2020). A bert-based approach with relation-aware attention for knowledge base question answering. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Malinka K, Peresíni M, Firc A, Hujnák O, Janus F (2023). On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree? In: *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education v. 1*, 47–53.

Manning CD, Raghavan P, Schütze H (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Mitchell T, Cohen W, Hruschka E, Talukdar P, Yang B, ..., Welling J (2018). Never-ending learning. *Communications of the ACM*, 61(5): 103–115. https://doi.org/10.1145/3191513

Newman D, Lau JH, Grieser K, Baldwin T (2010). Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Association for Computational Linguistics, Los Angeles, California.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, ..., Lowe R (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, ..., Riedel S (2019). Language models as knowledge bases? arXiv preprint: arXiv:1909.01066

Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741. https://doi.org/10.52202/075280-2338

Reimers N, Gurevych I (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint: arXiv:1908.10084

Ristoski P, Rosati J, Di Noia T, De Leone R, Paulheim H (2019). Rdf2vec: RDF graph embeddings and their applications. *Semantic Web*, 10(4): 721–752.

Robertson S, Zaragoza H (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Röder M, Both A, Hinneburg A (2015). Exploring the space of topic coherence measures. In: *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM)*, 399–408. ACM.

Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, ..., Rush AM (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint: arXiv:2110.08207

Suchanek FM, Kasneci G, Weikum G (2007). Yago: A core of semantic knowledge. In: *Proceedings of the 16th International Conference on World Wide Web*, 697–706.

Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, ..., Kenealy K (2024). Gemma: Open models based on Gemini Research and technology. arXiv preprint: arXiv:2403.08295

United Nations Economic Commission for Europe (UNECE) (2025). Generic statistical information model (GSIM) version 2.0: User guide. https://unece.org/. User Guide PDF. GSIM v2.0.

US Census Bureau (2025a). Census API user guide. https://www.census.gov/data/developers/

guidance/api-user-guide.html. Published January 16, 2025. Accessed September 1, 2025.

US Census Bureau, American Community Survey (2025b). American community survey (ACS). https://www.census.gov/programs-surveys/acs.html. Accessed September 1, 2025.

US Census Bureau, American Community Survey 1-Year Estimates (2023). American community survey 1-year estimates. https://api.census.gov/data/2023/acs/acs1. Accessed September 1. 2025.

US Census Bureau, American Community Survey 5-Year Estimates (2020). American community survey 5-year estimates. https://api.census.gov/data/2020/acs/acs5. Accessed September 1. 2025.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, ..., Polosukhin I (2017). Attention is all you need. Advances in neural information processing systems, 30.

Vrandečić D, Krötzsch M (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85. https://doi.org/10.1145/2629489

Wang J, Hu X, Hou W, Chen H, Zheng R, ..., Xie X (2023a). On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint: arXiv:2302.12095

Wang X, Wei J, Schuurmans D, Le QV, Chi EH, ..., Zhou D (2023b). Self-consistency improves chain of thought reasoning in language models. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR). ICLR*. 2023.

Wei J, Bosma M, Zhao VY, Guu K, Yu AW, ..., Le QV (2021). Finetuned language models are zero-shot learners. arXiv preprint: arXiv:2109.01652

Yang J, Jin H, Tang R, Han X, Feng Q, ..., Hu X (2024). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6): 1–32. https://doi.org/10.1145/3649506

Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q (2019). ERNIE: Enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A Korhonen, D Traum, L Màrquez, eds.), 1441–1451. Association for Computational Linguistics, Florence, Italy.