AutoDDG: Automated Dataset Description Generation using Large Language Models

Haoxiang Zhang, Yurong Liu, Wei-Lun (Allen) Hung, Aécio Santos, Juliana Freire
New York University
{haoxiang.zhang,yurong.liu,allen.hung,aecio.santos,juliana.freire}@nyu.edu

ABSTRACT

The proliferation of datasets across open data portals and enterprise data lakes presents an opportunity for deriving data-driven insights. However, widely-used dataset search systems rely on keyword searches over dataset metadata, including descriptions, to facilitate discovery. When these descriptions are incomplete, missing, or inconsistent with dataset contents, findability is severely hindered. In this paper, we address the problem of automatic dataset description generation: how to generate informative descriptions that enhance dataset discovery and support relevance assessment. We introduce AutoDDG, a framework for automated dataset description generation tailored for tabular data. To derive descriptions that are comprehensive, accurate, readable and concise, AutoDDG adopts a data-driven approach to summarize the contents of a dataset, and leverages LLMs to both enrich the summaries with semantic information and to derive human-readable descriptions. An important challenge for this problem is how to evaluate the effectiveness of methods for data description generation and the quality of the descriptions. We propose a multi-pronged evaluation strategy that: (1) measures the improvement in dataset retrieval within a dataset search engine, (2) compares generated descriptions to existing ones (when available), and (3) evaluates intrinsic quality metrics such as readability, faithfulness to the data, and conciseness. Additionally, we introduce two new benchmarks to support this evaluation. Our experimental results, using these benchmarks, demonstrate that AutoDDG generates high-quality, accurate descriptions and significantly improves dataset retrieval performance across diverse use cases.

PVLDB Reference Format:
Haoxiang Zhang, Yurong Liu, Wei-Lun (Allen) Hung, Aécio Santos, Juliana Freire. AutoDDG: Automated Dataset Description Generation using Large Language Models. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/VIDA-NYU/AutoDDG

1 INTRODUCTION

We have witnessed a proliferation of data portals and data lakes [17, 20, 24, 25, 32, 46, 49, 71] as more data is generated and made available. It is estimated that there are tens of millions of datasets on the web [48]. While these datasets present significant opportunities for data-driven insights, they must be easily findable and interpretable to maximize their utility [68].

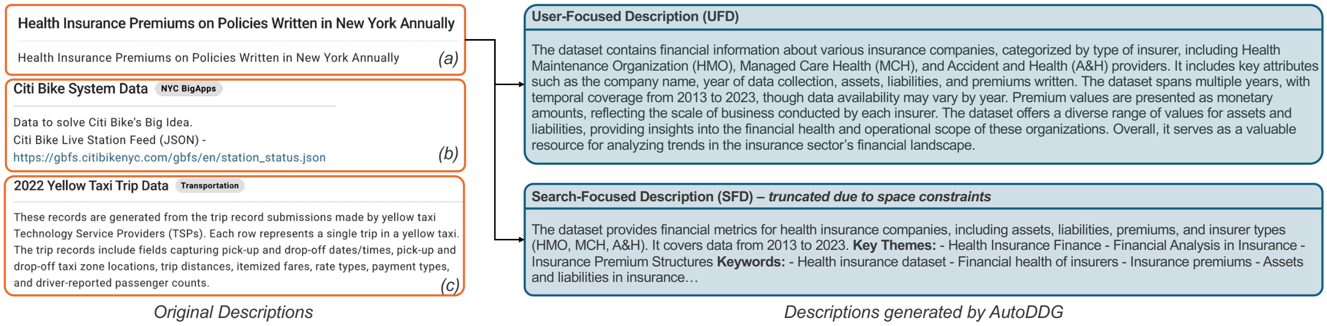
Many dataset search systems and infrastructure that powers data portals [4, 15, 58] follow the paradigm of web search engines: they rely on metadata—such as dataset names and descriptions–to build an inverted index for keyword-based queries. As a result, findability is directly influenced by the quality of dataset descriptions, particularly how well they convey dataset contents and align with users' information needs. For this reason, Google Dataset Search excludes from their index datasets discovered by the crawler which lack a description [3]. Open platforms like CKAN do not mandate descriptions but encourage them to improve search precision and recall, facilitate dataset understanding, and help users determine relevance [16]. However, many datasets are published with no descriptions or minimal ones. As a data point, consider the index of the Auctus dataset search engine [6]: 3,121 out of 23,520 datasets (13.2%) have no descriptions, while 2,346 datasets (approximately 10%) have descriptions with 10 words or fewer. Figure 1 illustrates issues with dataset descriptions in NYC Open Data [49]. Descriptions (a) and (b) fail to convey key characteristics of the dataset, such as their attributes, spatial coverage, or temporal extent. In some cases, descriptions are inconsistent with the actual contents; for example, description (c) implies the dataset includes taxi trips from 2022, whereas the actual data contains records that span multiple years.

Datasets published without detailed and accurate descriptions may be technically available but are effectively invisible. Koesten et al. [37] compare the problem of finding data today to the early days of the web, when people needed to know the URL of web pages or use manually created directories such as DMOZ to access content. Beyond findability, descriptions are also crucial for assessing dataset relevance. Just as web search users rely on snippets to decide which results to explore, dataset search users depend on descriptions to filter relevant datasets. This issue is even more pronounced for datasets, where downloading and inspecting large files is far costlier than simply opening a webpage.

Desiderata for Dataset Descriptions. Studies on information-seeking behavior and data discovery have revealed a gap between datasets that are available and those that users can effectively

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] The *Health Insurance* dataset:https://data.ny.gov/Economic-Development/Health-Insurance-Premiums-on-Policies-Written-in-N/xek8-zfrt/about_data
[2] The *Citi Bike* dataset: https://data.cityofnewyork.us/NYC-BigApps/Citi-Bike-System-Data/vsnr-94wk/about_data
[3] The *Yellow Taxi* dataset: https://data.cityofnewyork.us/Transportation/2022-Yellow-Taxi-Trip-Data/qp3b-zxtp/about_data


Health Insurance Premiums on Policies Written in New York Annually

Health Insurance Premiums on Policies Written in New York Annually

Citi Bike System Data

Data to solve Citi Bike's Big Idea.
Citi Bike Live Station Feed (JSON) -
https://gbfs.citibikenyc.com/gbfs/en/station_status.json

2022 Yellow Taxi Trip Data

These records are generated from the trip record submissions made by yellow taxi Technology Service Providers (TSPs). Each row represents a single trip in a yellow taxi. The trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off taxi zone locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

User-Focused Description (UFD)

The dataset contains financial information about various insurance companies, categorized by type of insurer, including Health Maintenance Organization (HMO), Managed Care Health (MCH), and Accident and Health (A&H) providers. It includes key attributes such as the company name, year of data collection, assets, liabilities, and premiums written. The dataset spans multiple years, with temporal coverage from 2013 to 2023, though data availability may vary by year. Premium values are presented as monetary amounts, reflecting the scale of business conducted by each insurer. The dataset offers a diverse range of values for assets and liabilities, providing insights into the financial health and operational scope of these organizations. Overall, it serves as a valuable resource for analyzing trends in the insurance sector's financial landscape.

Search-Focused Description (SFD) – ***truncated due to space constraints***

The dataset provides financial metrics for health insurance companies, including assets, liabilities, premiums, and insurer types (HMO, MCH, A&H). It covers data from 2013 to 2023. **Key Themes:** - Health Insurance Finance - Financial Analysis in Insurance - Insurance Premium Structures **Keywords:** - Health insurance dataset - Financial health of insurers - Insurance premiums - Assets and liabilities in insurance…

Original Descriptions

Descriptions generated by AutoDDG

Figure 1: Examples of datasets from NYC Open Data with inadequate ((a) *Health Insurance*[1] and (b) *Citi Bike*[2]) and inconsistent descriptions ((c) *Yellow Taxi*[3]) information. The *2022 Yellow Taxi Trip Data* dataset includes taxi trips that spans multiple years, not just 2022. Descriptions automatically generated by AutoDDG for dataset (a) are shown on the right. The complete SFD is given in the Appendix Table 10.

find [8]. These studies have also highlighted shortcomings in existing dataset metadata that hinder findability and relevance assessment [37, 51, 59]. Building on these findings, which we summarize in Section 2, we propose a desiderata for dataset descriptions:

Comprehensiveness and Alignment with Users' Information Needs: Widely-used search systems [4, 15, 58] often do not account for dataset contents, limiting users' ability to locate relevant data. Therefore, dataset descriptions should provide as much detail as possible about the contents, ensuring that they address users' information needs.

Faithfulness to Data: Descriptions must be accurate representations of the dataset contents.

Conciseness and Readability: Dataset descriptions serve a critical role in helping users assess relevance. They should be concise enough to facilitate quick evaluation yet detailed enough to convey meaningful insights.

Uniformity: While variability is expected in heterogeneous dataset collections, it is important to maintain a certain uniformity in the descriptions to facilitate comparison and relevance assessment.

There is an inherent tradeoff between conciseness and comprehensiveness: overly brief descriptions may fail to capture key aspects of the dataset, while excessively detailed ones can be difficult and time-consuming for users to process. Furthermore, while readability enhances users' ability to assess relevance, it is less critical for indexing and automated findability. Striking the right balance among these factors is essential for designing effective dataset descriptions. However, crafting high-quality descriptions manually is both difficult and time-consuming. This raises a critical question: *can dataset descriptions be generated automatically?* The ability to do so holds significant potential for improving dataset findability and enabling users to better navigate the vast amounts of data currently available.

LLMs for Automatic Description Generation. Large language models (LLMs) offer a promising avenue for generating dataset descriptions due to their ability to produce fluent, readable text and leverage broad world knowledge [5, 64]. These capabilities enable LLMs not only to generate coherent descriptions but also to enrich them with inferred semantic and contextual information. Prior work has demonstrated their effectiveness in semantic inference tasks for tabular data, including the ability to determine the semantic types of attributes and the table class [19, 34]. By connecting external knowledge that is not explicitly encoded in the data, LLMs can enhance dataset descriptions in ways that manual approaches might overlook.

However, using LLMs for this task presents several challenges. LLMs are designed for processing and generating textual data, whereas datasets are structured in tabular form. This mismatch raises questions about how best to represent tables in prompts—a challenge that remains an active area of research in table representation learning [62]. Moreover, LLMs operate within fixed context windows, limiting their ability to ingest and process large datasets in their entirety. As a result, we must rely on data samples, which introduces the risk of deriving incomplete or misleading descriptions, as a sample may fail to capture crucial global properties such as the dataset's spatial and temporal coverage.

Another key challenge is ensuring that generated descriptions align with the desiderata outlined earlier. Careful prompt design is essential to guide LLMs toward producing descriptions that are comprehensive, faithful to the data, and aligned with user information needs. Furthermore, LLMs are known to generate hallucinated or inaccurate content, making it critical to implement safeguards that ensure the descriptions remain grounded in the actual dataset.

Our Approach: AutoDDG. We introduce AutoDDG, an end-to-end system for automated tabular dataset description generation. Figure 2 outlines its key components. Given a tabular dataset, AutoDDG constructs a structured summary by combining two complementary profiles: a data-driven profile and a semantic profile (Section 3.2). The data-driven profile is generated using traditional data profiling techniques, capturing essential dataset characteristics such as attribute types, statistical summaries (e.g., value ranges), and distributions. The semantic profile, on the other hand, is derived with the assistance of an LLM and enriches the summary with contextual information, such as the dataset's topic. These profiles serve as input to the Description Generation Engine, which employs an LLM to generate textual descriptions (Section 3.3). The generated descriptions are then evaluated (Section 4.1) before being indexed by the dataset search engine.

By leveraging the data-driven profile, AutoDDG addresses two important challenges: 1) it ensures that the prompts reflect a global view of the data without relying on sampling, which may miss

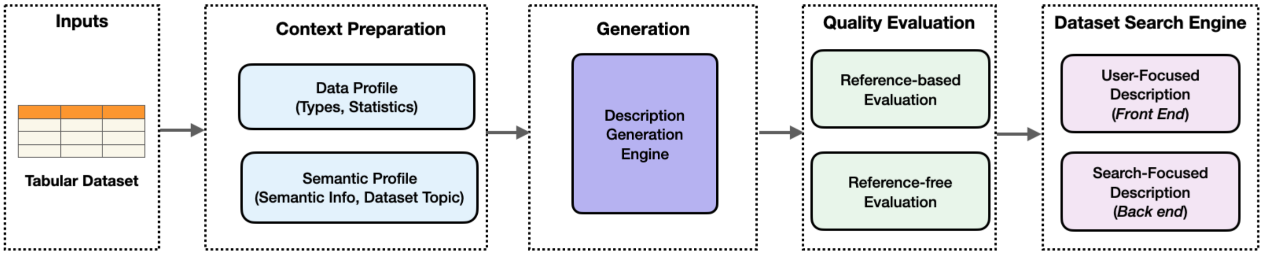

Inputs
Tabular Dataset
Context Preparation
Data Profile (Types, Statistics)
Semantic Profile (Semantic Info, Dataset Topic)
Generation
Description Generation Engine
Quality Evaluation
Reference-based Evaluation
Reference-free Evaluation
Dataset Search Engine
User-Focused Description (Front End)
Search-Focused Description (Back end)

Figure 2: *AutoDDG: a multi-stage framework for tabular dataset description generation.* The workflow begins with the *Context Preparation* stage, where the system processes the dataset to obtain dataset profile and semantic profile. In the *Generation* stage, the profiles are sent the *Description Generation Engine* powered by an LLM, which produces descriptions based on the processed input. The *Quality Evaluation* stage involves both *Reference-based* (if reference descriptions are available) and *Reference-free Evaluation*, enabling comprehensive quality assessment. Finally, the descriptions are utilized in a *Dataset Search Engine*, with *User-Focused Descriptions (UFD)* designed for front-end readability and *Search-Focused Descriptions (SFD)* optimized for back-end findability.

critical aspects; 2) it enhances the faithfulness of descriptions by grounding them in concrete dataset characteristics, reducing the risk of incomplete or misleading information.

The LLM-derived semantic profile further augments the descriptions with contextual information that may not be explicitly present in the data. As demonstrated in our experiments, this additional information improves dataset findability by enriching metadata with meaningful keywords and topical cues.

The profiling step plays a crucial role in extracting relevant metadata from both the dataset and the LLM. Therefore, the profilers should be designed to ensure alignment with the users' information needs. By using profilers, we also address a common challenge in dataset search engines: the inconsistency in the representation of metadata and descriptions for datasets from different providers and domains. Users "find it frustrating when metadata varies unpredictably across search results" [59]. By enforcing a consistent format, AutoDDG enhances the usability of dataset search tools.

As discussed above, balancing conciseness and comprehensiveness is a fundamental challenge in dataset description generation. To address this, we propose two types of description: User-Focused Descriptions (UFD), optimized for readability and clarity, provide users with a succinct yet informative overview of the dataset's content; and Search-Focused Descriptions (SFD), which incorporate dataset overviews, themes, and keyword-rich snippets to enhance search engine indexing (Section 3.3). As Figure 1 shows, UFDs offer concise, user-friendly summaries, while SFDs include additional metadata to improve discoverability. These two description types can be used in tandem and we evaluate to improve dataset search.

While it is possible to generate descriptions automatically, a crucial challenge is determining their quality. We propose a multi-pronged evaluation strategy that measures: 1) the improvement in dataset retrieval within a dataset search engine, 2) the similarity between the generated descriptions and existing ones (when available) to evaluate completeness and consistency, and (3) intrinsic quality metrics, using LLMs as a judge [22, 45, 57, 77] to evaluate readability, faithfulness to the data, and conciseness (Section 4.1). To conduct an evaluation, we need gold data that includes the relevance of a dataset for a given query. Since existing benchmarks only partially meet our needs, we introduce two new benchmarks specifically designed to support this evaluation (Section 4.2). In Section 5, we present a detailed experimental evaluation which shows that our approach is generates high-quality descriptions that lead to significant improvement in dataset retrieval. The results reinforce our design decisions and demonstrate the effectiveness of the different components of AutoDDG.

Summary of Contributions. To the best of our knowledge, this paper introduces the first comprehensive framework for systematically and automatically generating descriptions for tabular datasets, aiming to improve keyword-based dataset search. Our approach has the potential to significantly improve dataset findability by assisting dataset authors (creators) as well as managers of repositories and portals, in the creation of comprehensive and accurate descriptions. Our main contributions are as follows:

Problem Formulation We identify the Dataset Description Generation (DDG) problem for tabular datasets, outline the challenges involved, and the desiderata for effective descriptions.

Methods and Framework for Description Generation: We introduce AutoDDG, a framework that combines a data-driven approach and LLMs to generate informative descriptions for tabular data.

Evaluation Strategies: We propose a set of strategies to evaluate different quality aspects of descriptions, including *dataset retrieval evaluation*, *reference-based evaluation* and *reference-free evaluation*, to assess the quality of descriptions in the absence of reference descriptions.

Dataset Search Benchmarks: We construct two new benchmarks, ECIR-DDG and NTCIR-DDG, tailored to evaluate the quality of dataset descriptions for keyword-based search and retrieval tasks.

Experimental Evaluation: We report the results of an extensive experimental evaluation that validates our design decisions and demonstrates the effectiveness of AutoDDG at generating descriptions that improve dataset retrieval and that can assist users understand and assess dataset relevance.

2 RELATED WORK

Dataset Search. To better understand the behavior of users seeking structured data, Koesten et al [37] carried out a mixed-methods study, which encompasses interviews with users and analysis of search engine logs. One of the main findings of the interviews was that people often cannot find the data they need. The study also uncovered common strategies people use while searching for datasets, as well as *dataset attributes* that occur frequently in queries, including the category/topic of the dataset (e.g., transportation, business), geographic region covered by the data records, data granularity (e.g., spatial and temporal resolution); and how users determine the relevance of datasets, e.g., looking at summary statistics, understanding the semantics of data. Papenmeier et al. [51] conducted a survey with social science researchers who expressed their individual information needs for research data. They also found that "current systems fail to sufficiently support scientists in their data-seeking process" and mention that a key reason is that the metadata associated with datasets do not cover information that is sought by scientists. This underscores both the disconnect between the datasets a user needs and what datasets a user can actually find, as well as the importance of descriptions that properly describe the data and match users' information needs [8]. Sostek et al. [59] carried out a study that focused on the user experience with Google Dataset Search. They confirmed the metadata shortcomings identified in previous studies and acknowledge that there is a wide variability on metadata (description) quality, which depends on dataset provides and authors. They also identified an additional challenge: the inconsistency in how datasets are described leads to a higher mental load as users assess relevance and compare different datasets. We use the findings from these studies to guide the design of the data-driven and semantic profiles in AutoDDG. The automatic generation of descriptions can help address these challenges. By using the actual contents of the dataset to generate summaries and augmenting them with semantic information, we can derive high-quality descriptions. While it may not be possible to derive completely uniform descriptions as datasets contain different information, an automated process can be applied to all datasets and enforce a structure that makes the descriptions more uniform. For example, it is not possible to summarize the spatial extent of a dataset that does not contain spatial data, but for all spatial datasets, that information will be made available and represented in a similar fashion.

Discovery queries that go beyond keyword search have been proposed in which a dataset D is the query, and the results include datasets that are related to D, e.g., can be joined or concatenated, or are correlated [7, 35, 56, 78]. These queries are orthogonal to and can be combined with keyword search. The descriptions derived by AutoDDG have the potential to improve findability for both existing inverted-index based systems that are widely used, as well as for emerging dataset search engines like Auctus [6] that also support keyword-based search.

Table Understanding and Representation. Many models have been developed to enhance table understanding and facilitate various downstream tasks by semantically representing table content [18, 19, 26, 28, 29, 34, 38, 43, 61, 65, 67, 70, 72]. These areas include but not limited to table question answering, column type analysis and table type classification. For exmaple, TAPAS [26] proposes a weakly supervised table parsing model for question answering over tables. TURL [18] presents a novel framework for table understanding and representation learning, which can be fine-tuned for various tasks like entity linking and relation extraction. DODUO [61] introduces a column annotation framework based on pre-trained Transformer. Recently, large language models (LLMs) have begun to play a significant role in various data-related tasks. Korini and Bizer [38] use ChatGPT to address column type annotation with a two-step annotation pipeline, optimizing annotation efficiency. Chorus [34] focuses on data exploration tasks, such as table-class detection and column type annotation (CTA). ArcheType [19] establishes a new state-of-the-art for zero-shot CTA performance using LLMs. These studies demonstrate the growing application of foundation models to achieve state-of-the-art results in these areas. While these models advance semantic understanding in tables, they do not directly address the problem of generating dataset descriptions. Instead, these models focus on enhancing table comprehension and annotation, providing a foundation that could support enriched dataset descriptions but without the specific goal of generating comprehensive, human-readable summaries of entire datasets.

Text Generation from Table. Table-to-text generation models have attained significant attention due to their ability to transform structured data into coherent natural language text [11, 12, 23, 44, 53, 54, 57, 60, 63, 66, 76]. Among recent advancements, ReasTAP [76] introduces a novel table pre-training approach, which enhances models' reasoning capabilities through synthetic question-answering examples and demonstrates notable performance gains in producing logically faithful text in various downstream tasks. Additionally, the Multi-task Supervised Pre-training for Natural Language Generation (MVP) model [63] leverages multi-task supervised pre-training to excel in a broad range of natural language generation tasks, including knowledge-graph-to-text and data-to-text generation. However, these models are trained on datasets like Rotowire [69], WikiBio [39], and LogicNLG [10]. These datasets are designed for generating narrative text tailored to specific tasks such as sports game summaries and biographical sentences. While effective for generating natural language text from structured data, these models lack the ability to: (1) create descriptions for keyword-based search and high-level overviews; (2) handle large and heterogeneous datasets with complex structures. Moreover, they require training for a specific objective. In our scenario, however, we need to generate summaries that are general enough to accommodate a wide range of datasets and cater to different information needs, without requiring specific training for each dataset.

3 AUTODDG: AUTOMATING DATASET DESCRIPTION GENERATION

3.1 Problem Definition

Given a tabular dataset D with columns C and rows R, our task is to *automatically generate a descriptive summary* S_D that effectively captures the key characteristics of D, such as column names, data types, statistical properties, and semantically enriched information. To achieve this, we design a framework M that utilizes an LLM to generate the description S_D based on the prompt P and the context $X(D)$ derived from the dataset D:

$$S_D = M(P, X(D))$$

where P represents a template or set of instructions that guides the LLM to generate the desired type of description (e.g., user-focused or search-focused). $X(D)$ represents the context generated

Health Insurance Dataset

Number of Rows: 790
Number of Columns: 6
Columns:
- Name: Year
    - Data Types: Text, DateTime
    - Coverage: 2013 to 2022
    - Unique Values: 10
- Name: Liabilities
    - Data Types: Integer
    - Coverage: 0 to 2682301090.0
    - Unique Values: 757
    - ...

(other attributes are omitted due to space limitations)

Table 1: Example of data-driven profile.

by a Context Preparation module. Our problem involves designing effective methods for:

Context Preparation: Extracting and representing the dataset context $X(D)$ in a way that captures the heterogeneous characteristics of tabular data, fits within the input limitations of LLMs, and is aligned with users' information needs.

Description Generation: Crafting prompts P that guide the LLM to generate accurate and rich descriptions tailored to specific use cases, such as enhancing human readability or optimizing for search engines.

Quality Evaluation: Assessing the generated descriptions using appropriate metrics that evaluate their quality and impact on dataset findability, including both intrinsic measures (e.g., coherence, readability) and extrinsic measures (e.g., impact on search performance).

By formalizing the problem in this way, we set the foundation for developing an automated system that effectively generates dataset descriptions. Below, we describe how the components of AutoDDG address these challenges.

3.2 Extracting, Augmenting, and Representing Dataset Context

We focus on extracting and representing the context of tabular datasets to improve their interpretability and usability. To achieve this, we consider two types of context – data-driven and semantic – to create a summary of a dataset.

Data-Driven Profile. The data-driven context is derived directly from the dataset's contents, capturing structural and statistical properties that are commonly extracted by data profilers [1, 47, 52]. In our implementation, we use the Datamart Profiler [52]. As illustrated in Table 1, each table is profiled by analyzing all rows of each column to extract information such as attribute data types, value distributions, and uniqueness. These elements provide a global view of a dataset, summarizing its structure and content in a concise manner that can be effectively fed into LLMs for dataset description generation. Beyond statistical summaries, data-driven profiling can help domain-specific customization by extracting metadata that aligns with specific application needs.

Semantic Profile. The semantic profile goes beyond the dataset contents to capture contextual meaning that can enhance both human understanding and search precision, making dataset descriptions more informative. AutoDDG uses LLMs to enrich the dataset context with external knowledge that complements the data-driven profile and caters to users' information needs. This includes information about the dataset topic and usage, which improves interpretability and relevance [36]. Below, we describe the `Semantic Profiler` module (Algorithm 1) and the information it produces.

Algorithm 1 SemanticProfiler

```
Input: Tabular Dataset $D$, LLM model $M$, Number of Samples $sample\_size$
Output: Semantically enriched information for each column in $D$
Initialize list $semantic\_summary$ as empty
for each column $C_i$ in $D$ do
    $sample\_values$ = GetSample($C_i$, $sample\_size$)
    $semantic\_info$ = Prompt($M$, $C_i$, $sample\_values$)
    Create a human-readable summary $column\_summary$ based on the semantic information $semantic\_info$
    Append $column\_summary$ to $semantic\_summary$
end for
return $semantic\_summary$
```

3.2.1 *Structure-Defined Template Prompting.*

The `Semantic Profiler(SP)` uses a structured prompting approach to guide the LLM. For each column, it generates a prompt that includes the column name, sample values, and data type. The LLM responds with a JSON-formatted classification of the column based on predefined semantic categories. The structure-defined template and the complete prompt for semantic enrichment analysis are available in the Appendix Table 13. Examples of the output generated using the structure-defined template are shown in Table 2.

Example 1: *Year* column with values like 2018, 2020, 2023

Temporal:
- `isTemporal`: **True**
- `resolution`: **Year**

Spatial:
- `isSpatial`: **False**
- `resolution`:

Entity Type: Temporal Entity
Domain-Specific Types: General
Function/Usage Context: Aggregation Key

Example 2: *Liabilities* column with values like 137790801, 43992755, 599895

Temporal:
- `isTemporal`: **False**
- `resolution`:

Spatial:
- `isSpatial`: **False**
- `resolution`:

Entity Type: Monetary Value
Domain-Specific Types: Financial
Function/Usage Context: Measurement

Table 2: Examples of semantic profile.

Dataset Topic Generation Prompt

```
Using the dataset information provided, generate a concise
topic in 2-3 words that best describes the dataset’s
primary theme:
- Title: {title}
- Original Description: {original_description} (optional)
- Dataset Sample: {dataset_sample}
- Topic (2-3 words):
```

Table 3: Prompt for generating concise dataset topics based on the dataset title, description, and sample data.

The structured output is then serialized into descriptive sentences, and the summaries of all columns are concatenated to form the final output of the SP module. For example, the serialized semantic summary for example 1 is ***Year**: Represents temporal entity. Contains temporal data (resolution: Year). Domain-specific type: general. Function/Usage Context: Aggregation Key.* These enriched outputs provide a detailed understanding of each column’s semantic properties, enabling the generation of high-quality, contextually relevant dataset descriptions.

3.2.2 ***Dataset Topic Generation.*** To enhance dataset understanding and usability, we incorporate dataset topics into the *SP*. By leveraging large language models (LLMs), the process analyzes dataset metadata and samples to extract meaningful topics, which provide a high-level overview by capturing a dataset’s primary theme in 2-3 words. It involves two key steps: (1) Prompt Design, where a tailored prompt is constructed using the dataset’s title, original description (if available), and sample data to guide the LLM; (2) Topic Generation, where the LLM processes the prompt and generates a brief topic. Table 3 shows the prompt used by the Dataset Topic Generator.

3.2.3 *Discussion.* Note that the prompt design was guided by the findings of studies of information seeking requirements for dataset search (Section 2). Our goal is to obtain information that is aligned with how users formulate search queries. For example, the *SP* includes information about temporal and spatial attributes, as well as their resolution, as this was a common pattern observed by Koesten et al. [37]. The *SP* (and corresponding prompts) can be adapted to support other information needs and domains. It can also be extended to include other useful information, such as other semantic types of interest [19, 34].

3.3 Prompting LLM to Derive Description

AutoDDG uses LLMs to generate two types of dataset descriptions: User-Focused Descriptions (UFD) and Search-Focused Descriptions (SFD). For this task, we carefully design prompts to guide the LLM toward producing descriptions optimized either for presenting to dataset search engine users or for improving search relevance.

3.3.1 ***User-Focused Description (UFD).*** The User-Focused Description (UFD) is designed to provide a clear, concise, and accurate overview of the dataset, prioritizing human readability and presentation. This type of description works best for scenarios where the dataset needs to be communicated to users in a way that is easily understood, such as in reports, dashboards, or data catalogs aimed at human readers. Although its primary goal is readability, UFD can also be effective for search purposes, as it offers a well-structured overview containing key terms and concepts related to the dataset. The UFD is generated by prompting the LLM to describe the dataset based on dataset samples and the data-driven profile (Section 3.2). The prompt for generating a UFD and a typical output example are presented in Table 4. Note that the description includes information about the temporal extent and resolution, number of records, as well as a summary of the contents (e.g., the range of wind speed values) and an overview statement about the dataset and what it can be used for: it contains information about wind patterns and that it is useful for environmental studies.

User-Focused Description

Prompt: Answer the question using the following information. First, consider the dataset sample: *D_sample*. Additionally, the dataset profile is as follows: *D_profile*. Based on this profile, please add sentence(s) to enrich the dataset description. Furthermore, the semantic profile of the dataset columns is as follows: *D_semantic*. Based on this information, please add sentence(s) discussing the semantic profile in the description. Moreover, the dataset topic is: *D_topic*. Based on this topic, please add sentence(s) describing what this dataset can be used for. Based on the information above and the requirements, provide a dataset description in sentences. Use only natural, readable sentences without special formatting.

Example Output: This dataset contains wind speed and direction measurements from a specific time period in 2003. The data includes 4433 unique time stamps, with a temporal coverage of May 13 to June 12, 2003, at a resolution of minutes. The average wind speed ranges from 0 to 27.0, with a standard deviation ranging from 0 to 2.78. The average wind direction ranges from 0 to 338.0, with 16 unique values. The dataset provides a comprehensive view of wind patterns during this time period, making it suitable for environmental studies and research.

Table 4: Prompt and Example of User-Focused Description

3.3.2 ***Search-Focused Description (SFD).*** The Search-Focused Description (SFD) is optimized to enhance the findability of datasets in search engines and dataset portals. As illustrated in Figure 3, the process begins with a tabular dataset, which is processed by an LLM to generate an initial description and identify the dataset topic. By including a specific topic or area related to the dataset, the LLM can focus on expanding the description with relevant terms, concepts, synonyms, and keyword variations. This helps improve search engine indexing and retrieval performance. For example, suppose a user wishes to publish a dataset on "climate data" that contains detailed measurements across various regions. The system processes the dataset to generate an initial description and identifies "climate data" as the topic. The system then enhances the description by including climate-related keywords such as "temperature trends," "precipitation," "regional climate analysis," and "weather patterns." This process enables the SFD to incorporate a wide range of terms related to the topic, increasing the likelihood that users searching for climate-related datasets will find the dataset in question.

Table 5 shows a typical prompt for generating an SFD, which incorporates the dataset topic *D_topic* and an initial description *D_initial_description* as inputs, with an output structure defined by the SFD template. The result is a description that is densely

Search-Focused Description

Prompt: You are given a dataset about the topic *D_topic*, with the following initial description: *D_initial_description*. Please expand the description by including the exact topic. Additionally, add as many related concepts, synonyms, and relevant terms as possible based on the initial description and the topic. Unlike the initial description, which is focused on presentation and readability, the expanded description is intended to be indexed at backend of a dataset search engine to improve searchability. Therefore, focus less on readability and more on including all relevant terms related to the topic. Make sure to include any variations of the key terms and concepts that could help improve retrieval in search results. Please follow the structure of following example template: *Template*.

Example Output:

Dataset Overview: This dataset contains wind speed and direction measurements ... (An overview description based on UFD.)

Related Topics: - Climate/Weather Patterns - Renewable Energy - Wind Energy - Meteorology - Atmospheric Science - Environmental Monitoring - Sustainability - Climate Change - Weather Forecasting - Wind Turbine Performance - Energy Generation - Spatial Analysis

Concepts and Synonyms: - Wind Speed/Velocity - Wind Direction - Average Wind Speed - Standard Deviation - Climate Trends - Renewable Energy Sources - Energy Generation - Meteorological Data - Atmospheric Conditions - Weather Patterns - Climate Modeling - Weather Forecasting - Spatial Analysis - Temporal Data - Climate Change Impacts

Applications and Use Cases: - Analysis of wind patterns for renewable energy projects - Understanding climate trends and predicting wind behavior - Informing urban planning decisions to optimize wind energy harvesting - Developing weather forecasting models - Evaluating wind turbine performance and efficiency - Analyzing energy generation potential from wind resources - Studying the impact of climate change on wind patterns

Additional Context: - This dataset can be used to address questions such as "What are the typical wind patterns in a given region?" or "How does climate change affect wind behavior?" - It can be integrated with other datasets, such as climate models or energy consumption data, to provide a more comprehensive understanding of the relationship between wind patterns and energy generation. - The dataset's relevance extends to interdisciplinary applications, such as urban planning, sustainability, and environmental science.

(*additional information omitted due to space limitations*)

Table 5: Prompt and Example of Search-Focused Description

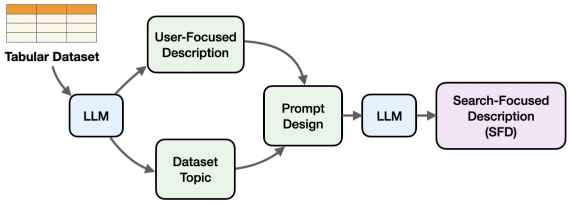


Figure 3: Workflow for generating Search-Focused Descriptions (SFD). The process begins with a tabular dataset, which is processed by a Large Language Model (LLM) to generate an initial description and identify the dataset topic. These components then feed into the prompt design stage to generate the final SFD, optimized for dataset findability and tailored to the provided dataset topic.

packed with terms relevant to the specified topic, which significantly improves the dataset's ranking in search results and makes it easier for users to discover the dataset when searching for related topics. A complete SFD example output and the output template are in the Appendix Table 10 and Table 14.

Discussion: UFDs and SFDs. While both UFD and SFD serve important roles in dataset description, they are optimized for different purposes. The UFD is crafted with the end user in mind, making the description easy to understand and presenting a well-rounded summary of the dataset. In contrast, the SFD is more technical and keyword-driven, focusing on enhancing search engine performance rather than prioritizing human readability. The generation of both types of descriptions allows AutoDDG to effectively serve both user-friendly and search-optimized needs.

There are different ways in which these descriptions can be generated. Here, we describe our initial approach that treats the SFD as an extension of the UFD. In the SFD, semantic information about the dataset's topic is reinforced to generate expanded keywords, concepts, and use cases, as illustrated in Table 5. Indexing SFD descriptions in search engines improves dataset findability by incorporating richer, more relevant terms. In future work, we plan to explore different strategies to combine the data-driven and LLM-derived information.

4 EVALUATION STRATEGY AND NEW BENCHMARKS

Our framework evaluates both the *intrinsic* quality of descriptions and their *extrinsic* impact on keyword-based retrieval. We describe the evaluation strategies as well as the metrics used. We also describe new benchmarks which we have tailored from existing resources to support the evaluation of dataset descriptions.

4.1 Evaluating Description Quality

We propose three complementary strategies to evaluate description quality: (1) *dataset retrieval evaluation*, which assesses the impact of descriptions on search effectiveness, (2) *reference-based evaluation*, which compares generated descriptions to existing ones using natural language generation metrics, and (3) *reference-free evaluation*, which leverages LLMs to assess descriptions without reference texts. The following subsections detail each approach.

Dataset Retrieval Evaluation. Evaluating the impact of generated descriptions on dataset retrieval performance is critical, as keyword-based search remains the primary application for dataset search engines. This evaluation specifically assesses the ability of description generation models to enhance the findability of datasets by improving search relevance and ranking—arguably the most important metric for practical use cases. To measure retrieval performance, we utilize the Normalized Discounted Cumulative Gain (NDCG@k) metric [30], a widely-adopted standard for ranking evaluation. NDCG@k measures the effectiveness of a ranking by comparing the ideal ranking of relevant items to the actual ranking produced by the system, up to position k. It accounts for both the relevance of the datasets and their positions in the result list, assigning higher scores when relevant datasets appear earlier.

In our evaluation, we integrate the generated descriptions into a keyword-based search engine and perform search queries representative of typical user behavior. By calculating NDCG@k scores for search results with and without the generated descriptions, we can quantify the improvement in dataset findability attributable to our method. This retrieval-based evaluation serves as the cornerstone of our assessment framework, providing a direct measure of the effectiveness of the descriptions in supporting dataset search engines. Unlike intrinsic evaluations, which focus on the quality of the descriptions themselves, this extrinsic evaluation measures their practical impact on search outcomes, highlighting their role in enabling better dataset findability.

Reference-Based Evaluation. For datasets that already have existing descriptions, a reference-based evaluation is applied by comparing the generated descriptions with the original ones. This comparison leverages traditional natural language generation (NLG) metrics such as METEOR [2] and ROUGE [41]. METEOR incorporates synonymy and stemming for a more flexible comparison, and ROUGE focuses on recall by evaluating overlapping units like n-grams and sequences. In addition, we use semantic similarity metrics like BERTScore [75], which employs contextual embeddings from BERT to capture the meaning and context of the text beyond surface-level word matches. These metrics provide a structured, quantitative evaluation of how well the generated description matches the reference in terms of wording, content, and meaning.

By comparing our generated descriptions to existing ones, we assess whether our automated approach produces descriptions of comparable quality. This is particularly important for data portals aiming to enhance or update their metadata, ensuring that automated descriptions meet the standards expected by users.

Reference-Free Evaluation. For datasets that do not have reference descriptions, reference-free evaluation becomes essential. In this scenario, we leverage large language models (LLMs) to evaluate the quality of the generated descriptions. Instead of relying on direct comparison with reference texts, LLM-based methods assess attributes such as coherence, relevance, clarity, and coverage of key dataset features [22, 45]. We prompt the LLM to rate the descriptions based on specific criteria. For example: "On a scale from 1 to 10, how well does the description capture the main characteristics of the dataset?" This provides a quantitative assessment of the descriptions' quality in the absence of reference texts.

However, it is important to acknowledge that LLM-based evaluations may introduce biases inherent in the models' training data. The models might favor certain styles or content, especially when evaluating outputs generated by the same model [50]. To mitigate this issue, we utilize cross-evaluation – for example, we use Llama to evaluate descriptions generated by GPT (and vice-versa). By employing different models for generation and evaluation, we reduce the likelihood of shared biases influencing the assessment. This reference-free evaluation approach complements traditional metrics by introducing semantic and contextual assessments, making it valuable for datasets without existing descriptions. It ensures that newly-published datasets have high-quality descriptions that effectively convey their content to potential users, thereby enhancing their findability.

4.2 New Dataset Retrieval Benchmarks

Challenges in Selecting Suitable Dataset Search Benchmarks. Identifying suitable evaluation benchmarks for keyword-based dataset search poses significant challenges due to the diverse nature of datasets and application scenarios. Since our goal is to generate descriptions that improve keyword-based dataset search, selecting appropriate benchmarks is critical for evaluating the performance of AutoDDG. We focus on tabular datasets in CSV format because they are widely used for representing structured data. Moreover, to simulate real-world scenarios, we aim to include CSV datasets that feature a large number of rows and columns, presenting a realistic and challenging evaluation environment. When selecting table search benchmarks, we consider the following criteria: (1) support for keyword-based queries; (2) inclusion of CSV datasets collected from Open Data Portals; and (3) availability of query relevance data, typically represented as triples of (CSV dataset, keyword query, relevance score). While there are several table search benchmarks[9, 14, 27, 31, 40, 42, 73, 74], only two meet these criteria: the ECIR [13] and the NTCIR [33] datasets.

However, both benchmarks require modifications before they can be used for our application scenario. In the ECIR benchmark, some datasets are labeled as relevant to a query, but closer (manual) examination reveals that they are not truly relevant. In the NTCIR benchmark, only 4.27% of the datasets are in CSV format. Building upon these two benchmarks, as we discuss below, we construct ECIR-DDG and NTCIR-DDG, tailored specifically for evaluating the effectiveness of description generation in keyword-based dataset search and retrieval.

ECIR-DDG Benchmark. The original ECIR benchmark focuses on tabular datasets from U.S. federal data sources. The queries are designed to reflect real-world information needs, with six distinct tasks covering various domains such as public health, economic data, and environmental statistics. Each query was manually created, and relevance judgments were obtained through crowdsourced annotations. We examined the original ECIR benchmark before constructing the new benchmark. For each query q, we first randomly sampled datasets D_{rel} that are labeled as relevant and datasets D_{irrel} that are labeled as irrelevant to q. We then manually reviewed dataset snippets—including the title, description, column names, and a sample of rows—to determine how many datasets in D_{rel} were truly relevant to the query q (true positives) and how many in D_{irrel} were truly irrelevant to q (true negatives). Our manual review revealed a true positive rate of 9.87%, with 23 out of 233 dataset-query pairs (D_{rel}, q) identified as valid. The true negative rate was 98.5%, with 65 out of 66 dataset-query pairs (D_{irrel}, q) confirmed as irrelevant.

Based on these results, we construct the new ECIR-DDG benchmark using the following steps. For datasets labeled as irrelevant to each query, we retain them as irrelevant candidates without modification. For relevant candidates, we issue the query through search engines such as Google Dataset Search and collect the top-ranked datasets. The collected datasets, along with their original titles and descriptions, are treated as relevant for the benchmark. The constructed ECIR-DDG benchmark includes 120 queries[4] with an average of 12 relevant datasets per query and 169 total datasets

[4] The original ECIR benchmark consists of 6 tasks, each with 20 queries.

Benchmark	Query	Rel Tabs/Query			Tabs/Query			Tabs/Bench	Rows/Tab			Cols/Tab		
		Avg	Min	Max	Avg	Min	Max		Avg	Min	Max	Avg	Min	Max
ECIR-DDG	120	12	8	16	169	76	549	1015	17977	4	183539	22	4	337
NTCIR-DDG	32	10	5	22	19	10	42	615	78630	3	2717008	14	2	74

Table 6: Statistics of the ECIR-DDG and NTCIR-DDG benchmarks. The table provides an overview of the number of queries, the average, minimum, and maximum number of relevant tables per query (Rel Tabs/Query), total tables per query (Tabs/Query), total tables per benchmark (Tabs/Bench), rows per table (Rows/Tab) and columns per table (Cols/Tab). These statistics highlight the diversity and scale of the datasets in the benchmarks.

per query (Table 6). The datasets vary in size, with average 17,977 rows and 22 columns. These statistics demonstrate the diversity and scale of the datasets in the benchmark.

NTCIR-DDG Benchmark. The original NTCIR benchmark is derived from Data.gov (US) and e-Stat (Japan), consisting of datasets across various domains such as climate, population, economy, and transportation. Queries in this benchmark were constructed from real user questions found in community Q&A forums and refined through crowdsourcing. The queries often include geographical, temporal, and numerical terms, reflecting common information needs in dataset search. Focusing on English topics, we filtered the dataset to include only CSV files, which account for 4.27% of the original NTCIR collection. To ensure sufficient data for meaningful evaluation, we selected queries with at least 5 relevant datasets and at least 10 total datasets (relevant and irrelevant combined). This filtering process resulted in a benchmark with 32 queries (Table 6) that meet these criteria, providing a targeted evaluation framework for keyword-based retrieval of CSV datasets.

5 EXPERIMENT EVALUATION

5.1 Experiment Overview

To evaluate our dataset description generation framework, we conduct experiments following the evaluation strategy introduced in Section 4.1 using multiple baselines and description models.

Models for Dataset Descriptions. We consider the following models for constructing descriptions:

(1) **Original**: the dataset descriptions provided with the datasets.

(2) **Header+Sample**: a simple method that concatenates column headers with sample values from each column.

Additionally, we evaluate two pre-trained language models by providing them the Header and Sample:

(3) **MVP**: the Multi-task Supervised Pre-training for Natural Language Generation (MVP) model [63] leverages multi-task training for improved text generation, including data-to-text generation.

(4) **ReasTAP**: a table pre-training model that incorporates reasoning skills to enhance table comprehension, including table-to-text generation [76].

Finally, we evaluate our proposed AutoDDG framework, which applies large language models (LLMs) to generate dataset descriptions automatically. We experiment with two variations of AutoDDG, and ablate them using different LLMs (GPT and Llama):

(5) **AutoDDG-UFD-GPT/Llama**: generates user-focused descriptions (UFDs) optimized for readability and clarity while maintaining relevance for search.

(6) **AutoDDG-SFD-GPT/Llama**: generates search-focused descriptions (SFDs) designed to improve dataset retrieval in keyword-based search engines.

Evaluation Metrics. We assess the generated descriptions using following metrics: (1) **Retrieval Metrics**: We measure retrieval effectiveness using NDCG scores from BM25 [55] and SPLADE [21]. BM25 evaluates keyword-based lexical matching, while SPLADE captures semantic relevance by expanding terms based on contextual meaning. (2) **Reference-Based Metrics**: We assess alignment with ground-truth descriptions using METEOR, ROUGE, and BERTScore, which capture phrase similarity, lexical overlap, and semantic consistency. (3) **Reference-Free Metrics**: We evaluate the descriptions' intrinsic quality using LLM-based scoring for Completeness, Conciseness, Readability, and Faithfulness. These metrics assess how well the descriptions convey relevant information while maintaining clarity and accuracy. Further details on these evaluation are provided in the corresponding sections.

Overview of Experimental Results. The radar chart in Figure 4 provides a comparative overview, illustrating how different methods perform across these dimensions. The results show that UFD excels in readability and conciseness, making it more suitable for user-facing applications, while SFD improves completeness and retrieval relevance, increasing dataset findability. By combining these strengths, AutoDDG achieves a balanced and superior performance compared to baseline methods (Original, Header+Sample,

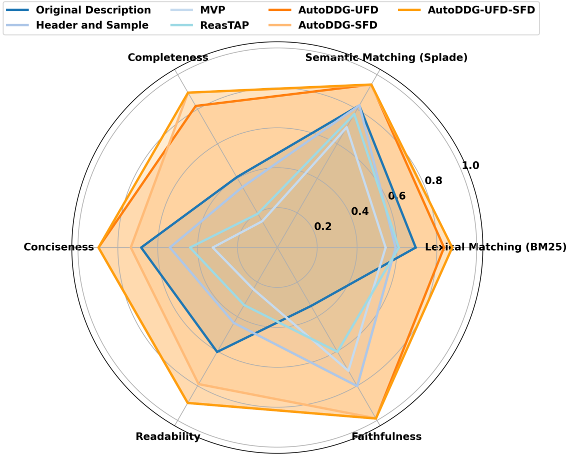


Figure 4: Radar chart comparing the performance of various dataset description methods across evaluation metrics. AutoDDG-UFD excels in Conciseness and Readability, while AutoDDG-SFD leads in Lexical Matching and Completeness. AutoDDG-UFD-SFD combines these strengths, outperforming baselines (Original, Header+Sample, MVP, ReasTAP). Metrics are scaled from 0 to 1.

MVP, and ReasTAP), demonstrating its effectiveness in generating high-quality descriptions for both human interpretation and machine-based search.

5.2 Retrieval Performance

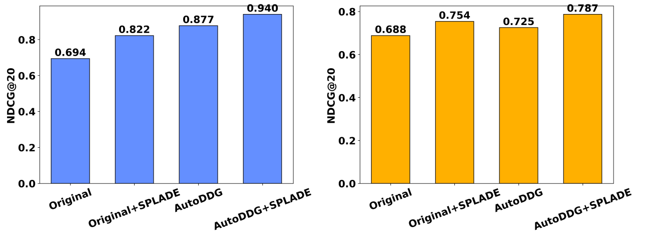
To assess the effectiveness of dataset descriptions in search tasks, we evaluate retrieval performance using **BM25** [55] and **SPLADE** [21]. BM25 focuses on lexical matching, ranking documents based on term frequency and inverse document frequency, while SPLADE learns sparse text representations by expanding query terms with semantically related words. For example, it may expand 'lunar' to include 'moon,' capturing broader context.

We experimented with both ECIR-DDG and NTCIR-DDG benchmarks. For the LLM models, we select cost-effective versions, GPT-4o-mini[5] and LLaMA-3.1-8B-Instruct[6], for our experiments. To evaluate retrieval performance, we created indices for both BM25 and SPLADE using the dataset descriptions generated by the models. For BM25, we created an inverted index, where terms from each dataset description were tokenized and stored along with their term frequencies and document frequencies.[7] During Evaluation, keyword-based queries were issued against this index, and BM25 ranked the dataset descriptions by calculating their relevance scores. For SPLADE, we created a sparse vector representation of each dataset description.[8] Each term in the description was assigned a weight based on its semantic importance, including terms absent from the text but inferred from the context (e.g., expanding "lunar" to "moon"). During evaluation, keyword-based queries were processed, and we ranked the dataset descriptions by calculating cosine similarity between the query's sparse representation and each document's sparse embedding.

BM25 Results. The results in Table 7 show that AutoDDG-SFD methods consistently outperforms UFD across both ECIR-DDG and NTCIR-DDG benchmarks, confirming that the effectiveness of SFDs. By tailoring descriptions to maximize keyword matching with query topics, SFDs improve search relevance. GPT leads to higher NDCG values than Llama on the ECIR-DDG benchmark, and Llama shows better performance on the NTCIR-DDG benchmark. This suggests that different LLMs exhibit varying strengths depending on the characteristics of the dataset and query topics.

The performance of UFD is competitive with SFD but, as expected, the focus on user readability limits its effectiveness in purely search-oriented metrics. Traditional baselines such as MVP and ReasTAP perform significantly worse than both UFD and SFD, highlighting their limitations in generating effective descriptions for retrieval tasks. Interestingly, original descriptions sometimes outperform these baselines, underscoring the importance of domain-specific context and the limitations of generic data-to-text approaches in keyword-based retrieval.

SPLADE Results. Our evaluation using SPLADE shows that semantic term expansion significantly enhances retrieval quality across all models. However, the adaptability of SFD to diverse query topics continues to set it apart as the best-performing method for retrieval tasks, particularly in benchmarks with more complex query characteristics. Figure 5 shows a comparison of the retrieval performance using the Original description and AutoDDG, with BM25 and SPLADE. The results consistently show that AutoDDG outperforms Original descriptions, with additional gains achieved when combined with SPLADE. This trend underscores the robustness of AutoDDG, even when combined with SPLADE, in enhancing retrieval effectiveness. A complete table of NDCG scores with SPADE is available in the Appendix Table 17.



(a) ECIR-DDG Benchmark

(b) NTCIR-DDG Benchmark

Figure 5: Summary plot of NDCG@20 scores for Original, Original+SPLADE, AutoDDG, and AutoDDG+SPLADE on ECIR-DDG and NTCIR-DDG benchmarks. This visualization highlights key improvements from AutoDDG and its combination with SPLADE.

5.3 Evaluation of Dataset Description Quality

In this section, we evaluate the quality of the generated dataset descriptions using both reference-based and reference-free methods (Section 4.1). These evaluation strategies offer insights into how well the descriptions meet human readability standards and perform in automated search and retrieval tasks. This is particularly useful for users looking to publish new datasets, as these evaluations can predict the performance of the generated descriptions both for front-end user interaction and back-end dataset retrieval. We apply following evaluation metrics:

METEOR: The METEOR score [2] combines precision, recall, synonymy, and stemming to measure the similarity between the generated and reference descriptions.

ROUGE: The ROUGE score [41] focuses on recall by measuring the overlap of n-grams, longest common subsequences, and skip-bigrams between the generated and reference texts.

BERTScore: BERTScore [75] utilizes pre-trained language models, such as BERT, to measure the similarity of embeddings between the generated text and the reference text. Unlike exact match metrics, BERTScore evaluates semantic similarity by comparing the contextualized representations of the words in both texts, providing a more meaningful assessment of whether the generated description conveys the same information as the reference.

LLM-Based Evaluations: LLM-based evaluations leverage large language models (LLMs) to assess the quality of generated descriptions, especially when reference descriptions are unavailable. Instead of direct comparison with reference texts, LLMs evaluate key attributes such as coherence, relevance, clarity, and coverage of dataset features [22, 45].

[5] https://platform.openai.com/docs/models

[6] https://deepinfra.com/meta-llama/Meta-Llama-3.1-8B-Instruct

[7] We used Okapi BM25 from this library: https://github.com/dorianbrown/rank_bm25

[8] We used SPLADE with FastEmbed: https://qdrant.github.io/fastembed/examples/SPLADE_with_FastEmbed/

	ECIR-DDG				NTCIR-DDG			
Model	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**
Original	0.666	0.658	0.668	0.694	0.439	0.575	0.662	0.688
Header+Sample	0.589	0.573	0.575	0.593	0.323	0.451	0.553	0.590
MVP	0.557	0.538	0.533	0.543	0.356	0.472	0.575	0.617
ReasTAP	0.638	0.598	0.599	0.609	0.406	0.485	0.583	0.627
AutoDDG-UFD-GPT	0.803	0.807	0.816	0.836	0.415	0.529	0.624	0.670
AutoDDG-UFD-Llama	0.739	0.741	0.766	0.788	0.425	0.533	0.631	0.669
AutoDDG-SFD-GPT	**0.846**	**0.849**	**0.857**	**0.877**	**0.512**	**0.613**	**0.688**	**0.725**
AutoDDG-SFD-Llama	0.763	0.764	0.784	0.809	0.437	0.567	0.651	0.689

Table 7: BM25. Comparison of NDCG scores across different description generation models on the ECIR-DDG and NTCIR-DDG benchmarks. Higher NDCG scores indicate better alignment between the generated descriptions and the relevance of retrieved results for keyword-based search. Bold values represent the highest scores for each metric, while underlined values indicate the second highest.

Model	METEOR	ROUGE	BERT Score
Original	-	-	-
Header+Sample	2.28	3.61	75.17
MVP	1.55	1.83	78.85
ReasTAP	6.48	8.82	79.99
AutoDDG-UFD-GPT	15.48	29.78	**82.24**
AutoDDG-UFD-Llama	**16.46**	30.08	82.26
AutoDDG-SFD-GPT	13.00	**34.50**	80.07
AutoDDG-SFD-Llama	12.27	33.85	79.17

Table 8: Evaluation of dataset description quality on reference-based metrics. Higher values indicate better performance. Bold represents the highest scores for each metric, while underlined values represent the second highest.

	Comp	Conc	Read	Faith
Model	**(G/L)**	**(G/L)**	**(G/L)**	**(G/L)**
Original	4.07/5.02	6.83/7.35	6.04/7.14	0.34/0.27
H+S	3.50/3.56	5.39/4.00	4.36/4.15	0.80/0.65
MVP	1.52/1.57	3.24/3.25	2.41/2.84	0.71/0.48
ReasTAP	1.96/1.91	4.38/5.81	3.33/4.85	0.60/0.50
UFD-GPT	8.19/8.57	**8.97/8.00**	**8.99/8.99**	0.99/0.90
UFD-Llama	7.72/8.43	8.51/7.86	8.63/8.50	0.96/0.92
SFD-GPT	**8.96/8.98**	7.35/5.97	7.90/7.00	**0.99**/0.93
SFD-Llama	8.95/8.89	6.24/5.56	7.19/6.79	0.97/**0.95**

Table 9: Evaluation of dataset description quality on reference-free metrics (Comp: Completeness, Conc: Conciseness, Read: Readability, Faith: Faithfulness) for all models. Scores are reported as GPT/Llama pairs. Completeness, Conciseness, and Readability are scaled from 0 to 10, while Faithfulness is scaled from 0 to 1. Higher values indicate better performance. Bold represents the highest scores for each metric, while underlined values indicate the second highest.

We follow the prompt design from G-Eval [45], incorporating task introduction, evaluation criteria, and evaluation steps. Additionally, we include example evaluations to guide the LLM in rating dataset descriptions. We focus on evaluation criteria ***completeness***, ***conciseness***, and ***readability***. They are chosen because they effectively assess the quality of dataset descriptions and predict their performance in different contexts. For instance, compared to search-focused descriptions (SFD), user-focused descriptions (UFD) are expected to score higher in conciseness and readability, but lower in completeness. In addition, we consider ***faithfulness*** evaluation to measure the extent to which the generated descriptions accurately reflect the underlying dataset content including dataset sample, dataset profile and semantic profile. We follow the prompt design in previous works [57, 77] to assess the truthfulness of dataset descriptions. The complete prompts for LLM-based evaluation can be found in Appendix Table 15 and Table 16. This approach provides a quantitative measure of the description's quality in the absence of references. However, it is important to account for potential biases inherent in LLMs, as they may favor certain styles or content [50]. To mitigate these biases, we employ cross-evaluation techniques using models with higher intelligence, GPT-4o and LLaMA-3.1-70B-Instruct, to evaluate the generated descriptions and report the average score from these two models.

Reference-Free Evaluation. In the reference-free evaluation (Table 9), AutoDDG models consistently outperform other methods. The AutoDDG-UFD-GPT model excels in *Conciseness* and *Readability*, confirming the expectation that user-focused descriptions (UFDs) prioritize clarity and readability. This reinforces our design decision of creating user-centered descriptions. Conversely, AutoDDG-SFD-GPT ranks highest in *Completeness*, as expected from search-focused descriptions (SFDs) that prioritize comprehensive data coverage, even at the expense of readability or conciseness. For the evaluation of *Faithfulness*, AutoDDG also outperforms other methods. Despite numerical differences, GPT and Llama evaluations show consistent trends across all reference-free metrics, further validating the reliability of these assessments. One interesting observation is that the *Original* descriptions provided with the datasets exhibit low faithfulness. This is understandable, as domain experts often include extra knowledge beyond what is explicitly contained in the dataset when creating descriptions, whereas generated descriptions rely solely on the dataset content. For example, a dataset containing statistics on wind speed, wind time, and wind direction may not include any information about the location of these measurements. Domain experts might add sentences about the location in the original description, but since this information is not present in dataset itself, the evaluation would consider these sentences unfaithful. This underscores the importance of having the dataset creators enrich automatically-generated descriptions to include

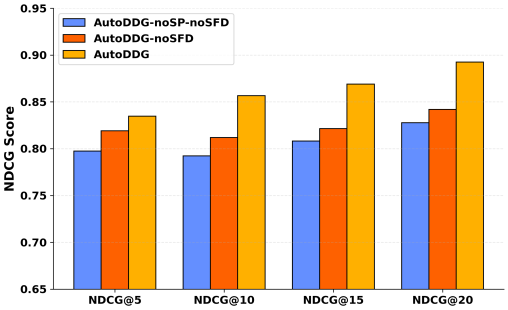


Figure 6: Comparison of NDCG scores across different settings of AutoDDG (noSP-noSFD, noSFD, and full configuration) on BM25. The results demonstrate the impact of enabling semantic profile (SP) and search-focused description (SFD) on retrieval performance.

metadata that cannot be automatically inferred, in particular, the provenance of the data.

Reference-Based Evaluation. The trends observed in the reference-based evaluation (Table 8) are consistent with the other results. The AutoDDG-UFD models perform exceptionally well in *BERT Score* and *METEOR*, reflecting better semantic alignment with reference descriptions. This suggests that UFDs, designed with human-centered presentation in mind, better capture the intent and meaning of the original data. On the other hand, AutoDDG-SFD models achieve superior scores in *ROUGE*, signaling a higher overlap with the reference text's exact terms and phrases. This aligns with the typical goal of search-focused descriptions to ensure high precision in matching key data points.

Summary. The findings from our experimental evaluation confirm our hypothesis: UFDs score higher in conciseness, readability, and semantic coherence (as seen in *BERT Score*), while SFDs perform better in completeness and exact overlap with reference descriptions. These insights suggest that while UFDs are more suitable for user engagement and semantic clarity, SFDs may be better for datasets where comprehensive and precise detail is paramount for search-related tasks. Therefore, using both description in a dataset search engine enables both a better user experience and search result quality.

5.4 Ablation Study

To assess the contributions of each module in AutoDDG, we conduct an ablation study by selectively disabling key components and analyzing their impact on retrieval performance. This allows us to isolate the effects of the Semantic Profile (SP) module and the Search-Focused Description (SFD) module. We compare the following variations: (1) **AutoDDG-noSP-noSFD**: Generates descriptions without SP and SFD, relying only on basic dataset samples and statistics. (2) **AutoDDG-noSFD**: Includes SP but omits SFD, enhancing descriptions with semantic insights while not optimizing for search retrieval. (3) **AutoDDG**: Includes both SP and SFD, leveraging semantic insights while optimizing descriptions for search retrieval.

Impact of Different Modules in AutoDDG. Figure 6 summarizes the results, measured by NDCG scores. The inclusion of the semantic profile (SP) module (AutoDDG-noSFD) improves retrieval compared to the baseline (AutoDDG-noSP-noSFD). Adding the search-focused description (SFD) module (AutoDDG) provides further gains, particularly when initial descriptions are concise, reinforcing its role in enhancing search relevance.

6 CONCLUSIONS AND FUTURE WORK

Effective dataset descriptions are essential for improving findability and assisting users in assessing dataset relevance. However, many datasets lack informative descriptions, limiting their discoverability and usability in search systems. In this paper, we introduced AutoDDG, an end-to-end framework for automated dataset description generation, designed to systematically address this challenge. AutoDDG combines data-driven profiling with LLM-powered semantic augmentation to generate high-quality descriptions that balance comprehensiveness, faithfulness, conciseness, and readability. We proposed a multi-pronged evaluation strategy for dataset descriptions that measures improvements in dataset retrieval, assesses alignment with existing descriptions, and employs LLM-based scoring for intrinsic quality evaluation. Recognizing the limitations of existing benchmarks, we introduced two new dataset search benchmarks, ECIR-DDG and NTCIR-DDG, to enable rigorous assessment.

Our experimental results demonstrate that AutoDDG significantly improves dataset retrieval performance, produces high-quality, accurate descriptions, and helps users better understand and assess datasets. Beyond its immediate impact on dataset search engines, our framework provides a scalable and systematic approach to metadata enhancement, benefiting data repositories, open data portals, and enterprise data lakes.

Limitations and Future Directions. While AutoDDG demonstrates strong performance for tabular datasets, an important avenue for future work is extending its applicability to a broader range of dataset types and content structures. In particular, we aim to explore multi-modal approaches that incorporate additional data modalities, such as images, to generate richer descriptions.

Our data-driven and semantic profilers were designed to extract key information identified as critical for dataset search [37, 51, 59]. Expanding their capabilities to support customization for diverse domains and use cases could further enhance their effectiveness, allowing domain-specific adaptations to better serve researchers and practitioners across different fields. Additionally, while our experiments relied on low-cost LLMs, an open question remains regarding the trade-offs between efficiency and performance when using larger, more powerful models. Future studies could investigate whether more advanced models significantly improve description quality or if lightweight alternatives suffice for most applications.

A critical challenge in LLM-generated descriptions is hallucination, where the model may introduce incorrect or misleading details. Developing robust detection and mitigation strategies to ensure descriptions remain faithful to the dataset's content and context is an important research direction.

Finally, while automatic description generation significantly improves dataset findability, the generated descriptions are necessarily incomplete. Exploring human-in-the-loop techniques, where users can enrich (e.g., add provenance and information on data collection methodology) or validate descriptions, presents an opportunity to further enhance accuracy, completeness, and trustworthiness, particularly in high-stakes domains such as healthcare and scientific research.

REFERENCES

[1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. *The VLDB Journal* 24 (2015), 557–581.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Omar Benjelloun, Shiyu Chen, and Natasha Noy. 2020. Google dataset search by the numbers. In *International Semantic Web Conference*. Springer, 667–682.

[4] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The World Wide Web Conference*. 1365–1375.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.

[7] Raul Castro Fernandez, Jisoo Min, Demitri Nava, and Samuel Madden. 2019. Lazo: A Cardinality-Based Method for Coupled Estimation of Jaccard Similarity and Containment. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 1190–1201. https://doi.org/10.1109/ICDE.2019.00109

[8] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. *The VLDB Journal* 29, 1 (2020), 251–272.

[9] Qiaosheng Chen, Weiqing Luo, Zixian Huang, Tengteng Lin, Xiaxia Wang, Ahmet Soylu, Basil Ell, Baifan Zhou, Evgeny Kharlamov, and Gong Cheng. 2024. ACORDAR 2.0: A Test Collection for Ad Hoc Dataset Retrieval with Densely Pooled Datasets and Question-Style Queries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 303–312.

[10] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. *arXiv preprint arXiv:2004.10404* (2020).

[11] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8635–8648.

[12] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2019. Few-shot NLG with pre-trained language model. *arXiv preprint arXiv:1904.09521* (2019).

[13] Zhiyu Chen, Haiyan Jia, Jeff Heflin, and Brian D Davison. 2020. Leveraging schema labels to enhance dataset search. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*. Springer, 267–280.

[14] Zhiyu Chen, Shuo Zhang, and Brian D Davison. 2021. WTR: A test collection for web table retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2514–2520.

[15] CKAN. 2024. https://ckan.org/.

[16] CKAN. 2024. CKAN User guide. https://docs.ckan.org/en/2.10/user-guide.html.

[17] dataverse [n.d.]. Dataverse Project. https://dataverse.org.

[18] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record* 51, 1 (2022), 33–40.

[19] Benjamin Feuer, Yurong Liu, Chinmay Hegde, and Juliana Freire. 2024. ArcheType: A Novel Framework for Open-Source Column Type Annotation Using Large Language Models. *Proc. VLDB Endow.* 17, 9 (May 2024), 2279–2292. https://doi.org/10.14778/3665844.3665857

[20] figshare [n.d.]. Figshare. https://info.figshare.com.

[21] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[22] Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383* (2024).

[23] Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1978–1988.

[24] Kathleen Gregory and Laura Koesten. 2022. *Human-centered data discovery*. Springer.

[25] James Hendler, Jeanne Holm, Chris Musialek, and George Thomas. 2012. US government linked open data: semantic. data. gov. *IEEE Intelligent Systems* 27, 03 (2012), 25–31.

[26] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4320–4333.

[27] Madelon Hulsebos, Çagatay Demiralp, and Paul Groth. 2023. Gittables: A large-scale corpus of relational tables. *Proceedings of the ACM on Management of Data* 1, 1 (2023), 1–17.

[28] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. 2019. Sherlock: A deep learning approach to semantic data type detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1500–1508.

[29] Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. Tabbie: Pretrained representations of tabular data. *arXiv preprint arXiv:2105.02584* (2021).

[30] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[31] Xingyu Ji, Aditya Parameswaran, and Madelon Hulsebos. 2024. TARGET: Benchmarking Table Retrieval for Generative Tasks. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

[32] Maxat Kassen. 2013. A promising phenomenon of open data: A case study of the Chicago open data project. *Government information quarterly* 30, 4 (2013), 508–513.

[33] Makoto P Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2021. A test collection for ad-hoc dataset retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2450–2456.

[34] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. 2024. Chorus: Foundation Models for Unified Data Discovery and Exploration. *Proc. VLDB Endow.* 17, 8 (April 2024), 2104–2114. https://doi.org/10.14778/3659437.3659461

[35] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationship-based Semantic Table Union Search. In *SIGMOD*.

[36] Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *International journal of human-computer studies* 135 (2020), 102367.

[37] Laura M Koesten, Emilia Kacprzak, Jenifer FA Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data: -a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1277–1289.

[38] Keti Korini and Christian Bizer. 2023. Column type annotation using chatgpt. *arXiv preprint arXiv:2306.00745* (2023).

[39] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771* (2016).

[40] Aristotelis Leventidis, Martin Pekár Christensen, Matteo Lissandrini, Laura Di Rocco, Katja Hose, and Renée J Miller. 2024. A Large Scale Test Corpus for Semantic Table Search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1142–1151.

[41] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[42] Tengteng Lin, Qiaosheng Chen, Gong Cheng, Ahmet Soylu, Basil Ell, Ruoqi Zhao, Qing Shi, Xiaxia Wang, Yu Gu, and Evgeny Kharlamov. 2022. ACORDAR: a test collection for ad hoc content-based (RDF) dataset retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2981–2991.

[43] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *International Conference on Learning Representations*.

[44] Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13415–13423.

[45] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[46] Fatemeh Nargesian, Erkang Zhu, Renée J Miller, Ken Q Pu, and Patricia C Arocena. 2019. Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment* 12, 12 (2019), 1986–1989.

[47] Felix Naumann. 2014. Data profiling revisited. *ACM SIGMOD Record* 42, 4 (2014), 40–49.

[48] Natasha Noy and Omar Benjelloun. 2020. An Analysis of Online Datasets Using Dataset Search. https://research.google/blog/an-analysis-of-online-datasets-using-dataset-search-published-in-part-as-a-dataset.

[49] NYC OpenData. 2024. https://opendata.cityofnewyork.us/.

[50] Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*

(2024).

[51] Andrea Papenmeier, Thomas Krämer, Tanja Friedrich, Daniel Hienert, and Dagmar Kern. 2021. Genuine information needs of social scientists looking for data. *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 292–302.

[52] profiler [n.d.]. datamart-profiler. https://pypi.org/project/datamart-profiler. Accessed on Jan 2025.

[53] Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6908–6915.

[54] Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics* 9 (2021), 510–527.

[55] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[56] Aécio Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation sketches for approximate join-correlation queries. In *Proceedings of the 2021 International Conference on Management of Data*. 1531–1544.

[57] Kwangwook Seo, Jinyoung Yeo, and Dongha Lee. 2024. Unveiling Implicit Table Knowledge with Question-Then-Pinpoint Reasoner for Insightful Table Summarization. *arXiv preprint arXiv:2406.12269* (2024).

[58] socrata [n.d.]. Socrata. https://open-source.socrata.com.

[59] Katrina Sostek, Daniel M. Russell, Nitesh Goyal, Tarfah Alrashed, Stella Dugall, and Natasha Noy. 2024. Discovering Datasets on the Web Scale: Challenges and Recommendations for Google Dataset Search. *Harvard Data Science Review* Special Issue 4 (apr 2 2024). https://hdsr.mitpress.mit.edu/pub/psnc8zsr.

[60] Yixuan Su, Zaiqiao Meng, Simon Baker, and Nigel Collier. 2021. Few-shot table-to-text generation with prototype memory. *arXiv preprint arXiv:2108.12516* (2021).

[61] Yoshihiko Suhara, Jinfeng Li, Yuliang Li, Dan Zhang, Çağatay Demiralp, Chen Chen, and Wang-Chiew Tan. 2022. Annotating columns with pre-trained language models. In *Proceedings of the 2022 International Conference on Management of Data*. 1493–1503.

[62] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 645–654.

[63] Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. MVP: Multi-task Supervised Pre-training for Natural Language Generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

[64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[65] Daheng Wang, Prashant Shiralkar, Colin Lockard, Binxuan Huang, Xin Luna Dong, and Meng Jiang. 2021. Tcn: Table convolutional network for web table interpretation. In *Proceedings of the Web Conference 2021*. 4020–4032.

[66] Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. 2021. Sketch and refine: Towards faithful and informative table-to-text generation. *arXiv preprint arXiv:2105.14778* (2021).

[67] Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. Tuta: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1780–1790.

[68] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[69] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052* (2017).

[70] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8413–8426.

[71] zenodo [n.d.]. Zenodo. https://zenodo.org.

[72] Dan Zhang, Yoshihiko Suhara, Jinfeng Li, Madelon Hulsebos, Çağatay Demiralp, and Wang-Chiew Tan. 2019. Sato: Contextual semantic type detection in tables. *arXiv preprint arXiv:1911.06311* (2019).

[73] Shuo Zhang and Krisztian Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 world wide web conference*. 1553–1562.

[74] Shuo Zhang and Krisztian Balog. 2021. Semantic table retrieval using keyword and table queries. *ACM Transactions on the Web (TWEB)* 15, 3 (2021), 1–33.

[75] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[76] Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir Radev. 2022. ReasTAP: Injecting Table Reasoning Skills During Pre-training via Synthetic Reasoning Examples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9006–9018.

[77] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. 2023. Investigating Table-to-Text Generation Capabilities of LLMs in Real-World Information Seeking Scenarios. *arXiv preprint arXiv:2305.14987* (2023).

[78] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J. Miller. 2019. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. Association for Computing Machinery, New York, NY, USA, 847–864. https://doi.org/10.1145/3299869.3300065

Search Focused Description (SFD) Example

Dataset Overview:

- The dataset provides comprehensive financial information pertaining to various health insurance companies, specifically focusing on different types of insurers, including Health Maintenance Organizations (HMO), Managed Care Health (MCH), and Acciden & Health (A&H) insurance providers. It includes pivotal financial attributes of these companies such as company name, year of data collection, total assets, liabilities, and premiums written. Spanning from 2013 to 2023, the dataset reflects a breadth of financial metrics, indicating the fiscal condition of these insurers and presents premium values categorized in monetary formats. This dataset aims to furnish insights into the financial landscape of the health insurance sector.

Key Themes or Topics:

- Health Insurance Finance
- Health Insurance Types (HMO, MCH, A&H)
- Financial Analysis in Insurance
- Risk Assessment in Health Insurance
- Insurance Premium Structures
- Temporal Changes in Health Insurance

Applications and Use Cases:

- Supports analysis for health insurance professionals and executives in evaluating financial health.
- Assists policymakers in understanding the economic implications of health insurance.
- Enables researchers to investigate trends and outcomes related to health insurance provisions.
- Useful for actuarial studies and financial modeling pertaining to health insurance.
- Facilitates comparative studies between different types of health insurance organizations.

Concepts and Synonyms:

- Health Insurance/Medical Insurance
- Financial Data/Financial Metrics
- Insurance Providers/Insurers
- Premium Income/Premiums Written
- Liability Assessment/Financial Liabilities
- Asset Valuation/Total Assets
- Risk Management/Insurance Risk
- Health Economics/Economic Aspects of Health Insurance
- HMO/Managed Care Organizations

Keywords and Themes:

- Health insurance dataset
- Financial health of insurers
- Insurance premiums
- Assets and liabilities in insurance
- Temporal financial data
- Insurance sector analysis
- HMO, MCH, A&H
- Company financial attributes
- Health insurance trends

Additional Context:

- The dataset is relevant for addressing current challenges and inquiries in the health insurance sector, especially regarding financial performance metrics during economic fluctuations.
- Its integration with health policy research and economic studies can provide a deeper understanding of how financial factors influence healthcare access and coverage.
- It can aid in developing frameworks for better risk management and investment strategies within the health insurance market.

Table 10: Search Focused Description (SFD) Example for a dataset on health insurance financials.

Category	Details
Temporal	• **isTemporal**: Does this column contain temporal information? Yes or No. • **resolution**: If Yes, specify the resolution (Year, Month, Day, Hour, etc.).
Spatial	• **isSpatial**: Does this column contain spatial information? Yes or No. • **resolution**: If Yes, specify the resolution (Country, State, City, Coordinates, etc.).
Entity Type	What kind of entity does the column describe? (e.g., Person, Location, Organization, Product).
Domain-Specific Types	What domain is this column from? (e.g., Financial, Healthcare, E-commerce, Climate, Demographic).
Function/Usage Context	How might the data be used? (e.g., Aggregation Key, Ranking/Scoring, Interaction Data, Measurement).

Table 11: Template for classifying column data into semantic types.

Response Example for Dataset Semantic Profiler

Response:

```
{
  "Domain-Specific Types": "General",
  "Entity Type": "Temporal Entity",
  "Function/Usage Context": "Aggregation Key",
  "Spatial": {
    "isSpatial": false,
    "resolution": ""
  },
  "Temporal": {
    "isTemporal": true,
    "resolution": "Year"
  }
}
```

Table 12: Response example showing a valid JSON output for classifying column data into semantic types.

Dataset Semantic Profiler Prompt

Instruction:

You are a dataset semantic analyzer. Based on the column name and sample values, classify the column into multiple semantic types.

Categories:

Please group the semantic types under the following categories:

- **Temporal**
- **Spatial**
- **Entity Type**
- **Data Format**
- **Domain-Specific Types**
- **Function/Usage Context**

Template Reference:

Following is the template: `{TEMPLATE}`

Rules:

(1) The output must be a valid JSON object that can be directly loaded by `json.loads`. Example response is `{RESPONSE_EXAMPLE}`.
(2) All keys from the template must be present in the response.
(3) All keys and string values must be enclosed in **double quotes**.
(4) There must be no trailing commas.
(5) Use **booleans (true/false)** and numbers without quotes.
(6) Do not include any additional information or context in the response.
(7) If you are unsure about a specific category, you can leave it as an empty string.

Dynamic Parameters:

- Column name: `{column_name}`
- Sample values: `{sample_values}`

Table 13: Prompt for the dataset semantic profiler.

Search-Focused Description Template

Dataset Overview:

- Please keep the exact initial description of the dataset as shown at the beginning of the prompt.

Key Themes or Topics:

- Central focus on a broad area of interest (e.g., urban planning, socio-economic factors, environmental analysis).
- Data spans multiple subtopics or related areas that contribute to a holistic understanding of the primary theme.

Example:

- theme1/topic1
- theme2/topic2
- theme3/topic3

Applications and Use Cases:

- Facilitates analysis for professionals, policymakers, researchers, or stakeholders.
- Useful for specific applications, such as planning, engineering, policy formulation, or statistical modeling.
- Enables insights into patterns, trends, and relationships relevant to the domain.

Example:

- application1/usecase1
- application2/usecase2
- application3/usecase3

Concepts and Synonyms:

- Includes related concepts, terms, and variations to ensure comprehensive coverage of the topic.
- Synonyms and alternative phrases improve searchability and retrieval effectiveness.

Example:

- concept1/synonym1
- concept2/synonym2
- concept3/synonym3

Keywords and Themes:

- Lists relevant keywords and themes for indexing, categorization, and enhancing discoverability.
- Keywords reflect the dataset's content, scope, and relevance to the domain.

Example:

- keyword1
- keyword2
- keyword3

Additional Context:

- Highlights the dataset's relevance to specific challenges or questions in the domain.
- May emphasize its value for interdisciplinary applications or integration with related datasets.

Example:

- context1
- context2
- context3

Table 14: Dataset description template outlining key themes, applications, concepts, keywords, and contextual relevance.

Dataset Description Evaluation Guidelines

Task: You will be given one tabular dataset description. Your task is to rate the description on three metrics.
Please make sure you read and understand these instructions carefully. Keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

1. Completeness (1-10):

- Evaluates how thoroughly the dataset description covers essential aspects such as the scope of data, query workloads, summary statistics, and possible tasks or applications.
- A high score indicates that the description provides a comprehensive overview, including details on dataset size, structure, fields, and potential use cases.

2. Conciseness (1-10):

- Measures the efficiency of the dataset description in conveying necessary information without redundancy.
- A high score indicates that the description is succinct, avoiding unnecessary details while employing semantic types (e.g., categories, entities) to streamline communication.

3. Readability (1-10):

- Evaluates the logical flow and readability of the dataset description.
- A high score suggests that the description progresses logically from one section to the next, creating a coherent and integrated narrative that facilitates understanding of the dataset.

Evaluation Steps:

- Read the dataset description carefully and identify the main topic and key points.
- Assign a score for each criterion on a scale of 1 to 10, where 1 is the lowest and 10 is the highest, based on the Evaluation Criteria.

Example Evaluations:

Example 1:

Description: The dataset provides information on alcohol-impaired driving deaths and occupant deaths across various states in the United States. It includes data for 51 states, detailing the number of alcohol-impaired driving deaths and occupant deaths, with values ranging from 0 to 3723 and 0 to 10406, respectively. Each entry also contains the state abbreviation and its geographical coordinates. The dataset is structured with categorical and numerical data types, focusing on traffic safety and casualty statistics. Key attributes include state names, death counts, and location coordinates, making it a valuable resource for analyzing traffic safety trends and issues related to impaired driving.

Evaluation Form (scores ONLY):

Completeness: 7, Conciseness: 9, Readability: 9

Example 2:

Description: The dataset provides a comprehensive overview of traffic safety statistics across various states in the United States, specifically focusing on alcohol-impaired driving deaths and occupant deaths. It includes data from 51 unique states, represented by their two-letter postal abbreviations, such as MA (Massachusetts), SD (South Dakota), AK (Alaska), MS (Mississippi), and ME (Maine). Each entry in the dataset captures critical information regarding the number of alcohol-impaired driving deaths and the total occupant deaths resulting from traffic incidents.

The column "Alcohol-Impaired Driving Deaths" is represented as an integer, indicating the number of fatalities attributed to alcohol impairment while driving. The dataset reveals a range of values, with the highest recorded number being 2367 deaths in Mississippi, highlighting the severity of the issue in certain regions. In contrast, states like Alaska report significantly lower figures, with only 205 alcohol-impaired driving deaths.

The "Occupant Deaths" column also consists of integer values, representing the total number of deaths among vehicle occupants, regardless of the cause. This data spans from 0 to 10406, with Mississippi again showing the highest number of occupant deaths at 6100, which raises concerns about overall traffic safety in the state.

Additionally, the dataset includes a "Location" column that provides geographical coordinates for each state, enhancing the spatial understanding of the data. The coordinates are formatted as latitude and longitude pairs, allowing for potential mapping and geographical analysis of traffic safety trends.

Overall, this dataset serves as a valuable resource for researchers, policymakers, and public safety advocates aiming to understand and address the impact of alcohol on driving safety across different states. It highlights the need for targeted interventions and policies to reduce alcohol-impaired driving incidents and improve occupant safety on the roads.

Evaluation Form (scores ONLY):

Completeness: 8, Conciseness: 7, Readability: 8

Final Evaluation Form:

Please provide scores for the given dataset description based on the Evaluation Criteria. Do not include any additional information or comments in your response.

Evaluation Form (scores ONLY):

Completeness: __, Conciseness: __, Readability: __

Table 15: Guidelines for evaluating dataset descriptions based on completeness, conciseness, and readability.

Dataset Description Faithfulness Evaluation Prompt

Task Description:

Your task is to evaluate whether the given dataset description faithfully represents the facts from the provided dataset sample, dataset profile, and semantic profile. For each sentence in the description, verify whether it accurately reflects the information from the inputs. Carefully check step-by-step for every sentence to determine if it is true or false. Use Example 1 as a reference and respond to Example 2.

Example 1

Dataset Sample:

```
Time Stamp, Average Wind Speed, Standard Deviation, Average Wind Direction
7/22/2003 1:40, 1.8, 1.66, 315
6/18/2003 10:50, 6.0, 1.44, 225
```

Dataset Profile:

- **Time Stamp:** Data is of type text, 13,576 unique values, semantic type: datetime.
- **Average Wind Speed:** Data is of type float, range: 0 to 30.2.
- **Standard Deviation:** Data is of type float, range: 0 to 3.05.
- **Average Wind Direction:** Data is of type integer, 16 unique values, range: 0 to 338.0.

Semantic Profile:

- **Time Stamp:** Represents temporal entity, resolution: Hour.
- **Average Wind Speed:** Represents measurement, domain: climate, function: measurement.
- **Standard Deviation:** Represents statistical measure, domain: general, function: measurement.
- **Average Wind Direction:** Represents measurement, domain: climate, function: measurement.

Data Topic: Wind Speed Data

Generated Description:

"This dataset contains 13,576 hourly records of wind speed and direction, with wind speed ranging from 0 to 30.2 m/s and wind direction values spanning from 0 to 360°. It includes detailed hourly timestamps as temporal entities, with wind speed and direction categorized as climate measurements. This dataset is useful for renewable energy planning, meteorological research, and understanding wind patterns."

Evaluation:

- **(Sentence 1):** "This dataset contains 13,576 hourly records of wind speed and direction, with wind speed ranging from 0 to 30.2 m/s and wind direction values spanning from 0 to 360°."
Explanation: The dataset profile specifies 13,576 records, the correct wind speed range (0–30.2 m/s), and the correct timestamp type. However, the wind direction range is incorrectly stated as 0–360° instead of 0–338°.
Verification: F
- **(Sentence 2):** "It includes detailed hourly timestamps as temporal entities, with wind speed and direction categorized as climate measurements."
Explanation: The semantic profile confirms that the timestamps are temporal entities, and wind speed and direction are categorized as climate measurements.
Verification: T
- **(Sentence 3):** "This dataset is useful for renewable energy planning, meteorological research, and understanding wind patterns."
Explanation: These applications are reasonable inferences based on the dataset's focus on wind speed and direction.
Verification: T

Example 2

Dataset Sample: `{dataset_sample}`
Dataset Profile: `{dataset_profile}`
Semantic Profile: `{semantic_profile}`
Data Topic: `{data_topic}`
Generated Description: `{dataset_description}`

Evaluation:

(Sentence 1): `{{sent1}}`
(Explanation): `{{explanation}}`
(Verification): T/F
...
(Sentence n): `{{sent n}}`
(Explanation): `{{explanation}}`
(Verification): T/F

Table 16: Prompt for evaluating dataset description faithfulness based on dataset sample, profile, and semantic profile.

	ECIR-DDG				NTCIR-DDG			
Model	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**
Original	0.833	0.826	0.820	0.822	0.614	0.698	0.734	0.754
Header+Sample	0.808	0.794	0.802	0.823	0.445	0.561	0.628	0.665
MVP	0.722	0.706	0.687	0.696	0.427	0.524	0.606	0.649
ReasTAP	0.799	0.768	0.758	0.772	0.526	0.607	0.659	0.695
AutoDDG-UFD-GPT	**0.889**	**0.908**	**0.934**	**0.943**	0.540	0.638	0.676	0.708
AutoDDG-UFD-Llama	0.876	0.897	0.922	0.933	0.569	0.651	0.706	0.735
AutoDDG-SFD-GPT	0.886	0.907	0.931	0.940	**0.644**	**0.712**	**0.773**	**0.787**
AutoDDG-SFD-Llama	0.877	0.892	0.921	0.932	0.615	0.709	0.745	0.772

Table 17: SPLADE: Comparison of NDCG scores across different description generation models on the ECIR-DDG and NTCIR-DDG benchmarks. Higher NDCG scores indicate better alignment between the generated descriptions and the relevance of retrieved results for keyword-based search. Bold values represent the highest scores for each metric, while underlined values indicate the second highest.