Fact and Fluency: A Neurosymbolic Approach to Modeling Data Science

Anonymous ACL submission

Abstract

Data scientists play a crucial role in linking data to information. They derive insights from data and then communicate those insights to stakeholders, providing them with a holistic view of the world the data describe. The typical data science process involves determining a user's information needs, mapping those needs onto data and analysis, and then communicating findings in a form that is understandable to that user. In this paper, we present an automated, neurosymbolic method, that mirrors the data science process, for generating factual documents from real-world datasets across a wide range of domains. In order to generate these factual documents, we utilize a suite of standard analytics to derive the facts and a large language model to express them fluently. Given a dataset in the form of a relational database, we use an analytics taxonomy to guide the production of plans and execute these plans to derive the necessary information from the database. This derived information is then handed off to a large language model to generate documents which are both factual and highly fluent. To evaluate this method, we generate 194 reports across 17 types of documents and 8 domains. We undertake a manual evaluation of these documents to determine the factual accuracy of generated reports.

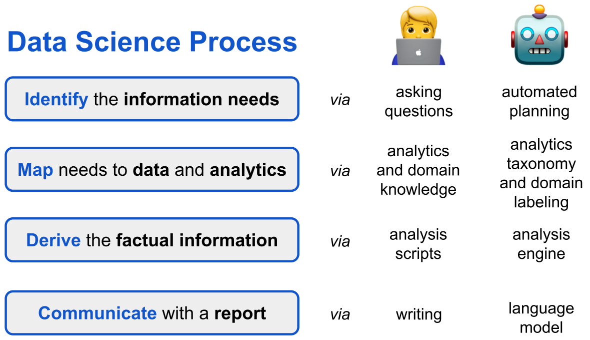

Figure 1: The data science process involves four key steps: identifying the information needs of a stakeholder, mapping these needs onto the available data and set of analytics which can be performed, using these analytics to derive the factual information from the data, and communicating the findings to the stakeholder with a written report. We map the ways a human data scientist carries out this process to how our automated, neurosymbolic approach does.

1 Introduction

There is a massive and ever growing amount of data from which people seek to derive novel insights. Data scientists engage in a process, shown in Figure 1, wherein they leverage their knowledge of the data, its domain, and analytic algorithms in order to produce contextualized information based on the needs of a stakeholder. To do this effectively, the data scientist must have a clear understanding of what this user's information needs are, how these needs map onto data and analysis, and be able to communicate their findings back to the stakeholder in a way that is understandable. A computational model of the data science process[1] would enable its automation and allow more people and organizations to derive insights from data, enabling better data driven reporting and decision making.

Recently, a taxonomy has been developed to model the analytic knowledge and processes that data scientists use to analyze data (Sterbentz et al., 2023). It provides a domain-independent taxonomy of analytic operations and how these map onto data. In order to effectively connect this taxonomy to real data, a domain labeling is created which maps tables and columns of relational data to entities, attributes, and the relationships between these entities. Each attribute is given a specific attribute type which specifies the analytic operations which can use it as input to derive meaningful information. A plan representation that is specified at the level of

[1] When we refer to data science, we specifically mean data analysis and communication, whereas the data ingestion and engineering steps are outside the scope of this approach.

this domain labeling was also introduced that details how to carry out this analysis with an analysis execution engine.

This formalization of data analytics enable us to model the information needs of a user by identifying common patterns of information requirements and mapping these to analytic plans. Executing these plans results in a discrete set of facts which must then be effectively communicated. One way to do this is via a document that organizes and summarizes these facts in an easily digestible manner. Recent autoregressive models such as GPT-4 (OpenAI, 2023) and Llama-2 (Touvron et al., 2023) have demonstrated an immense capacity for producing highly fluent and natural sounding text. By training on millions of documents, these models have learned a latent encoding of the structure of such documents. This property makes them an excellent tool for communicating information in the form of a document.

In this way, we seek to go from data to factual documents grounded by this data. Recent methods in this area tend to operate on tabular data which is embedded within documents (Puduppully et al., 2019; Yang et al., 2021) or small-scale tables (Suadaa et al., 2021). In contrast to these approaches, our goal is to produce documents whose information is grounded by large scale data stored within relational databases. The scale of data we seek to utilize means current end-to-end neural approaches (Sharma et al., 2022) are not sufficient due to their limited context sizes. Our approach is more akin to pipeline methods, such as (Kukich, 1983; McKeown, 1985; Reiter and Dale, 2000) in which the system first determines what to say by querying a knowledge or data source and then determines how best to communicate its findings.

In this paper, we seek to exploit the structural encodings of large language models in order to generate documents which are grounded by facts produced by a suite of analytics. In order to determine how best to leverage these structural encodings we first undertake an examination of the structure of five broad classes of documents and characterized the kinds of information they tend to produce by prompting a large language model to generate 125 reports and documents. This is discussed in more detail in Section 3. Such documents will be rife with confabulated information since they were not conditioned with any facts as part of the input. However, by identifying the types of information that are characteristic of these classes of documents, we can begin to identify what information should be provided to these models in order to ground their generation in fact. By providing these models with the facts and information relevant to the domain and data at hand, we hypothesize that these models can leverage their encodings of structure to produce language which is both highly fluent and adheres to the truth.

Section 4 presents a neurosymbolic method for modeling the data science process which enables the automatic generation of truthful documents grounded in real-world data. We leverage a taxonomy of data analytic processes and operations in order to perform complex analyses against relational data and derive key information. We define a set of document types that specify the kinds of information that would be useful to know about a particular entity. From these document definitions, plans which specify how to derive this information from the available data are produced and executed. The results are converted to natural language using a simple template-based technique which transforms the results to a sentence that a language model can use effectively. These results, along with instructions for generating the document, are joined into a final prompt which is passed to the language model. In order to ensure our approach is generalizable to a variety of domains, we apply this method to real-world datasets in 8 domains. In Section 5, we undertake a manual evaluation of documents generated with this method to determine how accurate the claims are in the reports and how well they adhere to the information that was derived from the database. We find that our approach generates reports whose claims are, on average, 83.5% factual, 2.5% refuted, and 13.8% confabulated. This indicates that our approach produces documents which strongly adhere to the underlying data while maintaining a high degree of fluency.

2 Related Work

Factual Language Generation Pretrained language models have shown a diverse range of adaptability across different tasks (Brown et al., 2020). They have been used to impressive degrees of effectiveness in law, medicine, education and other domains. (Petroni et al., 2019) showed that language models have the capacity of storing factual knowledge about entities. Inspired by this observation, a line of work focused on creating more effective probing techniques to elicit factual knowledge (Jiang et al., 2020; Shin et al., 2020; Zhong et al., 2021). Yet there have been challenges with language models being factual, or 'hallucinating', in a variety of settings including abstractive summarization (Maynez et al., 2020; Zhou et al., 2021), open-domain dialogue generation (Mielke et al., 2022; Roller et al., 2021), generative question answering (Li et al., 2021) and data-to-text generation (Dhingra et al., 2019). As mentioned in the introduction, our aim in this work is to utilize language models as a fluency engine. Instead of relying on it to correctly extract relevant facts, we prompt the model with all the factual information needed to complete the request.

Knowledge-Augmented Generation One of the key methods to promote factual generation by the language model is using an external knowledge base to augment the language model (Shuster et al., 2021). Such retrieval based methods started out as simple vector-space retrievers (Chen et al., 2017), and evolved into more end-to-end generation models where the retriever was jointly trained with the generator (Lewis et al., 2021; Izacard et al., 2022; Guu et al., 2020) which were shown to increase performance in downstream tasks. Knowledge graphs (Min et al., 2020), textual documents (Paranjape et al., 2021), pre-processed vectors (Verga et al., 2021), other language models (Shwartz et al., 2020) and search engines (Nakano et al., 2021) have all been used as external knowledge bases. Our work differs from the above ones in that instead of a learned or vector-space based retriever, we use an analytics engine which implements an analytics taxonomy and domain labeling to generate relevant information for each report type and prompt the language model with these generated facts.

Data-to-Text Generation A related task of data-to-text generation where the goal is to generate descriptions of structured data organized in tables has been studied for a long time (Kukich, 1983; Reiter and Dale, 2000). Traditionally, template based algorithms were used to build data-to-text systems (Oh and Rudnicky, 2000; Stent et al., 2004; Kondadadi et al., 2013), while recent approaches have adopted the planning then generation procedure (Su et al., 2021). However, our goals differ from the standard data-to-text generation goals in two key ways. First, our system focuses on end-to-end generation of full fledged reports given any dataset as opposed to description and summarization of a small number of highlighted cells in the table. Secondly, in the report generation process, we use the facts that are not present in the table, but which can be derived by taking advantage of the analytic capabilities afforded by the taxonomy.

3 Document Encodings of Large Language Models

In order to properly leverage the structure and fluency that language models provide, we first need to know what information is produced when the model generates a report. By identifying this, we can codify the information as an analytic plan template that our system can use to produce actionable plans to derive such information from the available data, and allow the language model to use this during its generation of a particular document type.

We seek to identify the information characteristics of 5 broad classes of documents: biographies, trends over time reports, comparison reports, analytic summaries, and performance reviews. For each of these classes, we generate 25 individual reports using GPT-4 with a temperature of 0. The 25 reports for each class were generated with a prompt constructed by taking a cross product of 5 specific instantiations of the report type across a variety of domains (to capture a broad range of contexts the document could be written for) and 5 instruction sets (to control for prompt variability affecting the final generation). A full listing of the document types and prompt templates which are combined to generate these documents can be found in Appendix A.

In total, we manually examine 125 documents generated by GPT-4 and analyze the information that was commonly communicated by them. In these reports, we identify factual claims being made and cluster them into characterizations of the information they present. These characterizations are summarized in Table 1.

Within each of the document classes, the structure of the generations tends to be fairly consistent, as does the kinds of information presented within them. We notice a slight exception to this for the *Analytic Summary* class, which has a greater variation in structure. We suspect this is due to the variety of metrics used in different domains being less consistent than, for example, reporting the date of birth, profession, and impact of a person in a *Biography*. As a result, when generating analytic summaries, it is important to have access to the analytics and metrics that are most appropriate for the domain and data. The relative consistency in

Document Type	Information Characteristics
Biography	Properties of the focus entity, what they are known for, early life and interests, how this interest led to their career, where and what they studied, achievements and awards, their impact on their professional field with examples, impact outside of their professional field, details of their personal life, and a statement of this person's overall importance.
Trends over Time	Introduction to the trend topic and region, what the general trend is in a certain timespan, how the trend has changed in recent years, factors causing the change in trend, explanation of the connection between these factors and the trend, subsets of the data for which the trends may be different, prediction of the trend in the future, effects of the trend continuing, and a conclusion with a summary of the trend, future expectations, and actions to take.
Comparison Report	Identification of the subject to be compared, statement of the comparison type being made, description of the subjects and how they relate to the comparison being made, the comparison of the subjects with regards to both the metrics and sub-metrics, statistics and evidence to back up these comparisons, and a concluding statement summarizing the similarities and differences of the subjects.
Analytic Summary	Introduction to the topic, why it is important, what factors may influence it, traits of the topic (e.g. facts, metrics, definitions, strengths and weaknesses), and a concluding statement summarizing what the topic is, measures of success or failure, and actions to be taken.
Performance Review	Identification of the subject being evaluated, their position/status, the period of time for the review, the report's purpose, review of the subject's performance in a variety of areas (for each of these a general statement on the level of competency was given along with supporting evidence and how they compare against similar entities), identification of the areas for improvement, a summarization of the subject's strengths and weaknesses, a recommendation for resolving the weakness, and prediction of how this subject will perform in the future.

Table 1: The five broad classes of documents we generate with the LLMs during our exploration of model encodings and the information that tend to be present. We find that there is a consistent structure for documents within each type, albeit less so for analytic summaries. This indicates that the model has a latent encoding of document structure and can be leveraged by conditioning the model with information grounded by real-world data, rather than producing confabulations with an unconditioned model.

structure and information presented indicates that this model has some latent encoding of document structure. We hypothesize that we can leverage this encoding by deriving similar kinds of information from a database and condition the model's generation with this information.

4 Factual Document Generation

The final output of the data science process we describe is a factual document which is grounded by the data and communicates key insights and information. In this section, we describe the neurosymbolic method which automates this process. The architecture for this approach is shown in Figure 2.

4.1 Analytics Taxonomy

For representing the analytic knowledge and processes that data scientists utilize, we adopt the analytics taxonomy presented by (Sterbentz et al., 2023). This taxonomy provides a model for specifying conceptual types for data and mapping these to relevant analyses via the taxonomy. A complementary domain labeling is also adopted which provides a means of specifying how the data is mapped to the conceptual types specified in the analytics taxonomy. These are used in tandem to model the way data scientists utilize their knowledge of analytic operations, underlying data types, and domain knowledge in order to determine what analyses can and should be performed with the data.

We also make use of the analytics engine presented in (Sterbentz et al., 2023) in order to execute analytic plans against data. This is used as a mechanism by which our method can derive factual information from data. This analytics engine maps analytic plans, specified at the domain level, to the equivalent SQL queries against the underlying data. The execution of these plans ultimately provides the raw information to be communicated as part of a document.

4.2 Information Requirements

In order for documents to be generated, the information to be conveyed must be identified and produced by analyzing the data. To accomplish this, we produce a set of analytic plan templates for each document type that we seek to generate. Together these plan templates constitute the *information requirements* for a document. These documents typically focus on a single entity instance, and the plan templates are filled with the specific entity being described by the document along with the attributes to be analyzed.

The information requirements for a document

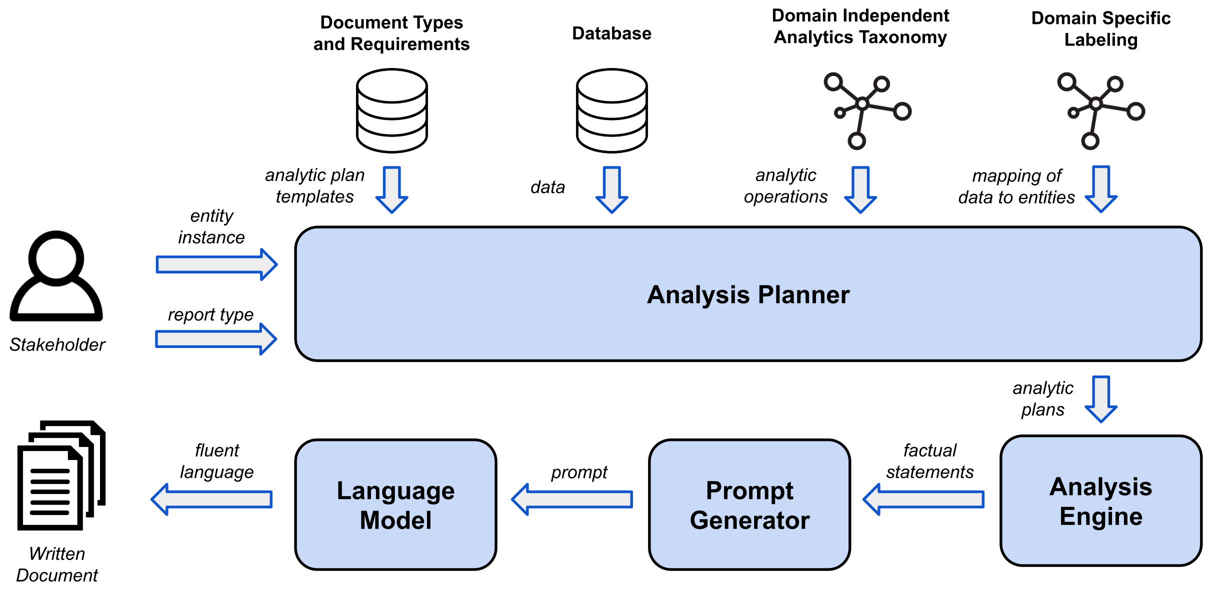

Figure 2: Our method leverages an analytic taxonomy and domain labeling to define the analytics that can be performed in terms of the entities in the database. A user inputs an entity instance for a desired document, for which the system will produce a corresponding set of analytics, execute it, produce factual statements based on the results, compile these into a prompt, and pass this to a language model to generate a report.

consist of properties about a particular entity instance, aggregations over its attributes, and time series of any event sequences. It is important that these reports convey relevant context with which to ground the information they provide about the focus entity. Doing so provides a more holistic view of the focus entity that supports a reader's understanding of the subject. To this end, we seek to provide aggregations over all entities of the same type, aggregations of similar entity instances according to a specific attribute, and rankings according to a specific attribute or aggregation. For example, the information requirements for a document focused on the income trends for a specific county in Illinois would include the name of the county, the resident population and median income for each year, and a calculation of the per capita income for the county. Each of these map to an individual plan produced from the corresponding analytic plan template. These plans are then executed with the analytics engine to produce the desired information to communicate with the language model.

4.3 Generation of Factual Statements

The analytic plans produced from the information requirements are ultimately converted to a SQL query against the underlying data, and the results of executing such queries are formatted as raw tables of data. The columns of these tables are dependent on the underlying data and can contain names which are difficult to interpret, such as those with abbreviations, underscores, and acronyms. Rather than relying on the language model to interpret these column names, we aim to provide the model with sufficient details about the analysis performed and the data being operated on. To achieve this, we leverage the knowledge provided by both the analytic taxonomy and the domain-specific labeling that specify simple language templates for analytic operations, descriptive names for entity attributes, and units for the attributes.

This linguistic knowledge is utilized during a recursive traversal of the analytic plan graph to generate human-readable sentences that describe what the results are. We structure these factual statements via two distinct generation modes and produce statement templates which the query results are slotted into.

4.3.1 Per Row

The first generation mode results in one factual statement per row of results. This mode has the benefit of being more human-readable, but is verbose and can be difficult to fit in the context size of some language models. The following is an example of two of the factual statements which were generated with this mode from the results of executing an analytic plan:

The state is AK and the count of unique incident *id grouped by state is 6.*

The state is AL and the count of unique incident id grouped by state is 61.

Note that each row from the results would be given its own statement, resulting in as many statements as there are rows of results.

4.3.2 Per Plan

The second generation mode results in one factual statement per plan. This mode produces statements that are more tabular, but contain descriptive names and units for the results, thus making the context suitable to fit within smaller context limits. But this comes at the cost of human readability, and all rows of results are ultimately compiled into a single statement. The following is an example of one factual statement generated with this mode that contains all results from executing an analytic plan:

The state and the count of unique incident id grouped by state is AK and 6, AL and 61, ...

4.4 Prompt Generation

Once a set of factual statements is generated, the system builds a prompt based on these facts and the type of document to produce. The final prompt is then passed to the language model to generate a fluent document for communicating the derived information. Our prompt consists of three key parts: (1) A description of the document to generate. Each document type has a short description that is used. (2) Instructions for the model on how the report should be generated. For example, "Generate a well-worded report." and "Use only the facts provided in the context." These instructions help steer the model towards fluency and factuality, respectively. (3) A set of factual statements.

5 Evaluation

In this section, we evaluate our approach for generating fluent and factual documents grounded by large scale data. We perform a manual evaluation of documents which are generated by our method.

5.1 Datasets and Domains

One of the key desiderata of our system is its ability to scale across a wide range of domains. To that end, we evaluate it over 9 datasets spanning 8 domains: healthcare, environmental sustainability, urban housing, criminal justice, education, legal and judicial, socioeconomic, and business. A summary of the datasets can be found in Table 4 in Appendix B. For each of these datasets, we generate a domain labeling that connects the data to the domain-independent analytics taxonomy. Based on the information characteristics identified in Section 3 and the data available in these datasets, we determine 17 document types to generate and the information that is to be derived for each. A detailed listing of the document types can be found in Appendix C.

5.2 Generating Analytic Documents

Each document type is focused on an entity. For example, the *Case Summary* document is focused on a specific instance of a legal *Case*. Three of the report types have only one instance of their entity of focus (*Handgun and Rifle Comparison*, *Summary of Shooting Incidents*, and *National Housing Price Trends*). For each of the other 14 document types, we randomly select 5 instances of the focus entity with which to perform the document generation.

Two recent large language models were used for generating the documents: GPT-4 (OpenAI, 2023) and StableBeluga-2 (Mukherjee et al., 2023). For the latter, we utilized a version[2] which had its 70 billion parameters quantized to 4-bits following the method in (Frantar et al., 2023). Note that via some brief experimentation with prompts, there were slight wording differences in the instructions provided to these models. For GPT-4, we found that instructing it to "Generate a 500 word report..." constrained it to generate brief, but factual documents. For StableBeluga-2, we observed that instructing it to "Generate a well-worded report..." reduced confabulation, while still generating fluent and factual documents. Additionally, there were some technical limitations with fitting the prompts into the 4K context window of StableBeluga-2. As a result, for StableBeluga-2 all of the *per row* prompts and some of the *per plan* prompts were too long for its 4K context window.

In the end, we generated 194 documents for evaluation, with 146 of them being produced by GPT-4 and 48 from StableBeluga-2. Some examples of documents generated with this method are shown in Appendix E.

5.3 Human Evaluation of Factuality

The key metric for evaluating the documents produced by our method is the percentage of claims

[2] https://huggingface.co/TheBloke/StableBeluga2-70B-GPTQ

in the generated documents supported by the factual statements provided as input by the analytics engine. We opted to ignore the fluency of the outputs as this is a highly subjective feature of the text and makes comparisons against other outputs difficult to measure. To determine the percentage of claims supported by the input factual statements, we perform manual evaluation. This process was carried out in two steps: claim identification and claim classification.

5.3.1 Claim Identification

We define a claim as any assertion of truth involving some retrieval or analytic processing of information from the data. Interstitial writing that provides a transition between content, titles, and broad introductions to the report are not considered claims. When identifying statements in the document as claims, we ignore ambiguous claims. That is, if the context of the claim directly impacts the truthfulness of the claim, we do not consider this a claim. For example, within the sentence "The temperature fluctuated greatly between 78 and 90," there would be two claims: that the min temperature was 78 and the max was 90. However, the statement that the temperature fluctuated greatly requires a subjective judgment of what it means to fluctuate greatly according to some metric. Therefore, this would not be marked as a claim. Using these definitions, we manually examine each generated document and identify the claims being made.

5.3.2 Claim Classification

Once a claim has been identified, we need to determine which of the following mutually exclusive categories this claim belongs to based on the factual statements provided to the language model.

Factual: The claim is directly stated by a fact in the context, or the claim can be directly inferred from the facts in the context via an unambiguous analytic process.

Refuted: The claim is directly refuted by a statement in the context, or the claim could be directly inferred from the facts in the context via an unambiguous analytic process, but is incorrect.

Confabulation: The claim is not directly pulled from the facts in the prompt, or is not inferred from the facts in the prompt.

Note that the *Confabulation* category includes subjective statements or conclusion (e.g. "We should be doing...", "This should be done...", etc.) and definitional statements that are not provided to the language model in the context. It is important to note that not all confabulations are bad. If the model were to define standard deviation in the document, this can be a useful piece of information. However, we do not want to rely on the model to present this, especially if these definitions are inadequate or imprecise. Rather, this is an indication that we should seek to present this definition directly to the model. Furthermore, although this is related to the term *hallucination*, widely used in the literature, we intentionally avoid using the term ourselves because our definition of confabulation differs from the multiple ways hallucination has been defined already.

		Factual	Refuted	Confab.
Gen. Mode	**Per Plan**	0.84	0.03	0.13
	Per Row	0.83	0.01	0.16
Model	**GPT-4**	0.84	0.01	0.15
	SB-2	0.84	0.07	0.09

Table 2: Scores in the table are percentage of total claims. Note that SB-2 is StableBeluga-2. Both models had similar factual accuracy, but GPT-4 had a greater tendency to confabulate, while StableBeluga-2 had a greater tendency to produce erroneous statements.

5.4 Results

In this section, we present some key quantitative and qualitative findings from our manual evaluation of the reports generated by GPT-4 and StableBeluga-2. Figure 3 shows the percentage of claims in the generated reports which were factual, refuted, or confabulated. It is broken down by domain to show the results for each of the 8 domains separately. The results for each domain includes multiple types of reports as well as reports generated by each of the two language models. For all domains we can see a high percentage of factual claims ranging from 69.6% for the criminal justice to 92.3% for healthcare. Exact values can be found in Table 6 in Appendix D.

Majority factual statements As shown in Table 2, more than eighty percent of claims made in the reports by either GPT-4 or StableBeluga-2 are factual. Although the fraction of factual claims are slightly higher for StableBeluga-2, we found that reports generated by GPT-4 contained more information as indicated by the 31.44 average number

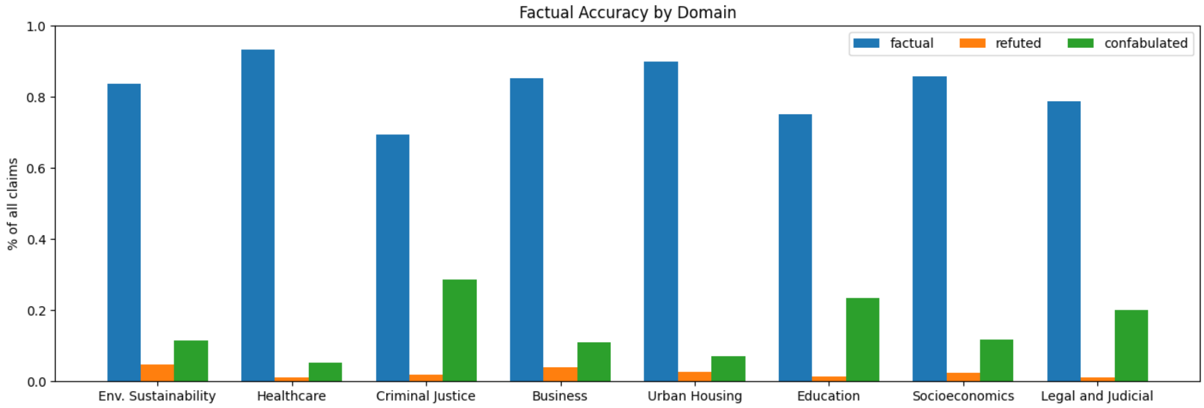

Figure 3: A breakdown of the percentage of claims which are factual, refuted, or confabulated for each of the 8 domains. These results are an aggregation of reports generated by both models (GPT-4 and StableBeluga-2).

of claims made by GPT-4 per document vs 14.94 by StableBeluga-2. This discrepancy is likely due to the limited context size of StableBeluga-2 preventing it from generating documents for which large amounts of information was derived from the data. Regardless of this, both models demonstrated a strong ability to adhere to the given set of factual statements when writing the documents.

Useful Claims It is difficult to objectively evaluate each claim made in the reports on their usefulness. However, qualitatively, we found more instances of vacuous factual claims in StableBeluga-2's reports. For instance, in a national rent trends report, it states *"When looking at the average monthly rent by region name, we can see that some areas are more expensive than others."* This indicates that, particularly from the GPT-4 model, statements made in the generated report are not only factual, but also tend to present insightful information.

End of Reports We found that often GPT-4 tended to end its reports by trying to frame the report in a larger context. This often led to confabulation. For instance, in a report about wildfires in California, it concludes with *"The state's wildfire situation is generally more severe than the national average, highlighting the need for effective fire management strategies in California."* Statements like these indicate that it may be useful to provide the model with high level concluding statements as part of the input context in addition to the facts derived via data analysis.

Above and Beyond GPT-4 had a tendency to format data nicely (i.e. 22:57:00 to 10:57 PM, 10:51:00 hours to 10 hours and 51 minutes), while StableBeluga-2 tended to parrot the input back more closely. At the moment, the documents contain examples of the model performing aggregations, comparisons, and inferences on its own. While the formatting is desirable, our goal is to move away from the model-based analysis by seeking to provide these analytic statements and use the model as an engine for generating fluent language only. It is left as future work to determine how best to prevent the models from making these ungrounded inferences.

6 Conclusion and Future Work

In this paper, we present a neurosymbolic method for modeling the data science process. This method generates factual documents based on information derived from real-world datasets across a wide range of domains. This method mirrors the one utilized by data scientists to derive and communicate meaningful information from data. A suite of analytics was used to derive the information, and a large language model was used to express them in fluent language. We validate this approach by generating and then manually examine 194 documents coming from 17 document types and 8 domains for factual correctness. We find that this method results in just 2.5% of claims in the generated documents being refuted by the data.

In the future, we aim to build an automated planner that is capable of utilizing the data directly when making decisions about what information is most meaningful to derive and communicate. We also intend to develop automated methods for fact checking the documents with the goal of using this as a feedback signal for a learning algorithm that we can use to improve the generation.

Limitations

A limitation to our document generation approach is that, in the system's current state, a set of information requirements must be specified for each new type of document able to be generated by the system. For example, adding a "salesperson performance review" document class would mean specifying a new set of plan templates to fulfill the information goals of this report (e.g., plan templates that retrieve total sales for that salesperson, a comparison of those sales with the trend of their past sales, a comparison to other similar salespersons, etc.). Our immediate next steps on improving this system include the development of more generalized information requirements that scale across many report types and a planner that can automating determine which information requirements to use based on a user's utterance.

Another decision made in our document generation method that makes it somewhat limited is our reliance on the LLM to encode the structure of the report type rather than controlling for this ourselves. While we do consider this somewhat desireable, as it prevents us from having to define an explicit structure for each new type of report we wish the system to generate, it also means that we have little control over the document structure and as such, the quality and coherence of the structure depends on how well it is encoded in the LLM.

Ethics Statement

In this work we present a system that automates the process of generating documents from data, using a taxonomy of analytic operations to aid in guiding what information gets presented in the report. Of the 9 datasets on which our system was evaluated, 8 of them are publicly available and the other is scheduled to be publicly available in the coming months.

Additionally, while the aim of our system is to ground automatically generated reports in fact, particularly as compared to an unguided LLM, the system does not preclude the generation of reports from factually flawed data. As such, it is contingent on the system's user to verify accuracy of the data fed to the system.

For reproducibility purposes, we plan to publicly release our code repository in the near future.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Center for Homeland Defense and Security. 2023. Shooting incidents at k-12 schools (jan 1970-jun 2022) - chds school shooting safety compendium.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Illinois State Board of Education. 2022. 2022 report card public data set.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know?

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Steven Horng, Roger Mark, and Massachusetts Institute of Technology [MIT]. 2023. Mimic-iv-ed demo v2.2.

Ravikumar Kondadadi, Blake Howald, and Frank Schilder. 2013. A statistical nlg framework for aggregated planning and realization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1406–1415.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 942–947, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Kathleen R McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial intelligence*, 27(1):1–41.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2020. Knowledge guided text retrieval and reading for open domain question answering.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Alice H. Oh and Alexander I. Rudnicky. 2000. Stochastic language generation for spoken dialogue systems. In *ANLP-NAACL 2000 Workshop: Conversational Systems*.

OpenAI. 2023. Gpt-4 technical report.

Andrew Paley, Andong L Li Zhao, Harper Pack, Sergio Servantez, Rachel F Adler, Marko Sterbentz, Adam Pah, David Schwartz, Cameron Barrie, Alexander Einarsson, et al. 2021. From data to information: automating data science to explore the us court system. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 119–128.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D. Manning. 2021. Hindsight: Posterior-guided training of retrievers for improved open-ended generation.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems (Studies in Natural Language Processing)*. Cambridge: Cambridge University Press.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Mandar Sharma, Ajay Gogineni, and Naren Ramakrishnan. 2022. Innovations in neural data-to-text generation. *arXiv preprint arXiv:2207.12571*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Karen C. Short. 2022. Spatial wildfire occurrence data for the united states, 1992-2020.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86.

Marko Sterbentz, Cameron Barrie, Donna Hooshmand, Shubham Shahi, Abhratanu Dutta, Harper Pack, Andong Li Zhao, Andrew Paley, Alexander Einarsson, and Kristian Hammond. 2023. Lightweight knowledge representations for automating data analysis. *Preprint*.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

United States Environmental Protection Agency. 2015. Air Data.

U.S. Bureau of Economic Analysis. 2022. Personal income in cook county, il.

U.S. Bureau of Labor Statistics. 2023. Unemployed persons in cook county, il.

U.S. Census Bureau. 2022a. Estimate of median household income for cook county, il.

U.S. Census Bureau. 2022b. Estimate of people age 0-17 in poverty in cook county, il.

U.S. Census Bureau. 2022c. Estimate of people of all ages in poverty in cook county, il.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural memoryover symbolic knowledge. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 3678–3691.

Yang Yang, Juan Cao, Yujun Wen, and Pengzhou Zhang. 2021. Table to text generation with accurate content copying. *Scientific reports*, 11(1):22750.

Yelp Inc. 2023. Yelp open dataset.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zillow Group, Inc. 2023. Housing data - zillow research.

A Prompts for Document Encoding Exploration

When examining the information characteristics of documents generated by a large language model, 125 individual prompts were used to create the same number of documents. The specific instances of each of the five document types can be found in Table 3. Each of these were slotted into to replace {report} for each the following prompt templates:

- Generate a 500 word {report}.
- Within 500 words, please write a {report}.
- Please write a 500 word {report}.
- Give me a {report}using a maximum of 500 words.
- Please compose a {report}. Limit the report to 500 words.

B Datasets

Table 4 lists the 9 datasets with brief descriptions, in the 8 domains across which we evaluated our system.

Air Data The Air Data (United States Environmental Protection Agency, 2015) provides a yearly overview of Air Quality Index (AQI) measurements, which assess the overall air quality by considering various air pollutants in a specific geographic region, such as counties or core-based statistical areas (CBSA). These summaries offer a combination of qualitative information, like the number of days with "good" air quality, and quantitative data, such as the median AQI value. The availability of these summary values can differ from one area to another due to variations in monitoring stations for different pollutants.

Wirdfire Occurrence The Wildfire Data (Short, 2022) is a comprehensive collection of information regarding wildfires in the United States spanning from 1992 to 2020. This dataset comprises

Document Type	Dataset
Biography	Biography of a fictional actor
	Biography of a fictional politician
	Biography of a fictional academic
	Biography of a fictional company or organization
	Biography of a fictional athlete
Analytic Summary	Summary of a patient stay at a hospital
	Summary of political contributions for a political party
	Summary of fundraising performance for a politician
	Analysis of the correlation between academic expenditures and student performance
	Summary of air quality and pollutants for a city
Comparison Report	Housing price comparison report for two metropolitan regions
	Report comparing the air quality comparison for two counties
	Report comparing the statistics associated with the usage of handguns vs rifles used in school shootings
	Report comparing two judges
	Report comparing the poverty levels of two counties
Trends over Time	National housing price trends report
	Housing price trends for a metropolitan region
	Academic performance trends for an individual student
	Wildfire occurrence trends in the United States
	Air quality trends for a county in the United States
Performance Review	Performance report for a salesperson
	Performance report for a doctor
	Performance report for a business
	Performance report for a community of businesses
	Progress report for an individual student

Table 3: A listing of each document instance type that was slotted into the prompt templates. Note that many of these document instance types correspond to the 17 document types we tested our method with in Section 5.

2.3 million records that are geographically referenced and represents 180 million acres of land that were consumed by wildfires over the course of 29 years. Additionally, it contains unique identifiers that enable the connection of individual wildfire data points to larger fire perimeter datasets and operational situation reports, providing a holistic view of the wildfire incidents.

MIMIC-IV-ED-Demo The Medical Information Mart for Intensive Care (MIMIC-IV-ED) database (Johnson et al., 2023) is a comprehensive repository of critical care data that includes information from more than 40,000 patients. The data has been deidentified to remove any patient identifiers in compliance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision, ensuring patient privacy. The data in MIMIC-IV-ED is derived from patients who were admitted to intensive care units at the Beth Israel Deaconess Medical Center (BIDMC).

Zillow Observed Rent Index The Zillow Observed Rent Index (Zillow Group, Inc., 2023) is a rental price index designed to accurately represent the entire rental housing market, rather than just the properties currently listed for rent. This index assigns dollar values by averaging the rents of homes and apartments that fall within the 40th to 60th percentile range for a specific region. It covers national, metropolitan, county, city, and zip code levels, with appropriate weighting to ensure it reflects the overall rental housing stock.

School Shooting Incidents This dataset (Center for Homeland Defense and Security, 2023) encompasses information regarding shooting incidents that have occurred from the start of 1970 up to June 2022. These incidents encompass a broad range of situations, including instances where firearms were displayed, discharged, or where bullets struck school property, irrespective of the number of casualties, the time of day, or the day of the week.

Illinois Report Card This dataset (Illinois State Board of Education, 2022) is an annual report published by the Illinois State Board of Education. It provides information about the progress of the state, individual schools, and districts in achieving various educational objectives, as well as school funding information at the state and federal level. The report offers a comprehensive overview of student and school performance to assist families and communities in understanding and aiding their local schools.

SCALES SCALES-OKN (Paley et al., 2021) utilizes two primary datasets: PACER, which serves as the official source for electronic federal judicial records, and the Federal Judicial Center's (FJC) database containing information about federally appointed judges. The SCALES-OKN dataset incorporates specific docket reports from PACER, encompassing ten years of docket data from Northern Illinois district courts between 2007 and 2016, as well as docket reports from all district courts for the year 2016. Additionally, it includes a variety of judge-related metadata, such as birthdate, gender, race/ethnicity, appointment history, appointing parties, educational background, and professional career details.

Income Disparity The Personal Income by County, Metro, and Other Areas report by the U.S. Bureau of Economic Analysis (U.S. Bureau of Economic Analysis, 2022; U.S. Bureau of Labor Statistics, 2023; U.S. Census Bureau, 2022b,a,c) provides information about the earnings of individuals residing in specific geographic regions, such as counties and metropolitan areas. This data reflects the income received by these residents or on their behalf, offering estimates based on their place of residence.

Yelp Open Dataset The Yelp Open Dataset (Yelp Inc., 2023) is a collection of Yelp's data, including reviews, information about businesses, and user data. It's available for personal, educational, and academic use. This dataset was compiled by Yelp and at the time of collection included 5,996,996 reviews, details about 188,593 businesses, and 280,992 pictures from the Yelp platform.

C Generated Document Types

As part of our evaluation, we generate 17 different kinds of documents across 8 domains: healthcare, environmental sustainability, urban housing, criminal justice, education, legal and judicial, socioeconomic, and business. A summary of all 17 document types that were generated as part of the evaluation of our neurosymbolic method are shown in Table 5.

C.1 Healthcare

Within this domain, we generate two unique kinds of reports: Patient Visit Report and Subject Report.

Patient Visit Report This report provides a summary of a single visit to the emergency department. Features which are reported on include: the visit duration, the arrival transport, patient triage information, their diagnosis, the number of vital signs which were taken during the visit, and times series of vital signs for heartrate, temperature, oxygen saturation, systolic blood pressure, and diastolic blood pressure.

Subject Report This report provides a summary of all visits made to the emergency department for a single patient. Features which are reported on include: personal characteristics of the subject, the number of stays they've had in the emergency department, and time series of their stays which denotes the stay id, intake time, duration of the stay, the acuity level, and diagnosis for each of their stays.

C.2 Urban Housing

Within this domain, we generate two unique kinds of reports: MSA Housing Price Trends and National Housing Price Trends.

MSA Housing Price Trends This report provides a summary of housing price trends for a particular metropolitan region of the United States. Features which are reported on include: the name of the region, its size ranked against other metropolitan areas, the region type, the most recent rent price reported for this area, a complete time series of the rent over the years, and the percent difference between this area's rent and the national average.

National Housing Price Trends

This report provides a summary of overall housing price trends in the United States. Features which are reported on include: the most recent national rent price average, a time series of the average rent across the country, the three regions with the highest average rents, and the three regions with the lowest average rents.

C.3 Criminal Justice

Within this domain, we generate three unique kinds of reports: Summary of Shooting Incidents, Shooting Incident Report, and Handgun and Rifle Comparison.

Summary of Shooting Incidents This report provides a summary of all gun-related incidents on school grounds for a single county, district, or state. Features which are reported on include: the total number of incidents reported, the number of incidents involving a handgun, the number of incidents involving a rifle, the number of shootings reported for each state, the number of shootings reported for each school level, and the number of shootings reported for each time period.

Domain	Dataset	Description
Environmental Sustainability	Air Data	The Air Data (United States Environmental Protection Agency, 2015) summarizes annual Air Quality Index (AQI) values for geographic areas, including qualitative measures and statistics, with availability varying due to monitoring station coverage.
	Wildfire Occurrence	The Wildfire Data (Short, 2022) provides 2.3 million geo-referenced records on U.S. wildfires from 1992 to 2020, covering 180 million burned acres with key identifiers for data linkage.
Healthcare	MIMIC-IV-ED-Demo	MIMIC-IV-ED (Johnson et al., 2023) is a deidentified critical care database from BIDMC with 40,000+ patient records, organized modularly for easy access to diverse data sources while complying with HIPAA Safe Harbor.
Urban Housing	Zillow Observed Rent Index	Zillow Observed Rent Index (Zillow Group, Inc., 2023) is a representative dollar-denominated rental index, calculated from listed rents in the 40th to 60th percentile for all housing types in various regions.
Criminal Justice	School Shooting Incidents	The dataset (Center for Homeland Defense and Security, 2023) covers publicly available data on shooting incidents from 1970 to June 2022, including any instance of gun brandishing, firing, or bullets hitting school property, regardless of outcomes or timing.
Education	Illinois Report Card	The Illinois Report Card (Illinois State Board of Education, 2022), issued by the Illinois State Board of Education, provides annual educational progress data for the state, schools, and districts.
Legal and Judicial	SCALES	The SCALES dataset (Paley et al., 2021) combines data from PACER, including ten years of docket reports (2007-2016) from Northern Illinois district courts and 2016 district court reports, with the Federal Judicial Center's judge metadata.
Socioeconomic	Income Disparity	The U.S. Bureau of Economic Analysis' report (U.S. Bureau of Economic Analysis, 2022; U.S. Bureau of Labor Statistics, 2023; U.S. Census Bureau, 2022b,a,c) captures personal income data for various regions, showing income received by residents in those areas based on their place of residence.
Business	Yelp Open Dataset	The Yelp Open Dataset (Yelp Inc., 2023) comprises 5.9 million reviews, 188,593 businesses, and 280,992 pictures, provided by Yelp for personal, educational, and academic purposes.

Table 4: Datasets used for the evaluation which specify the underlying data used for each of the 8 domains we tested with.

Shooting Incident Report This report provides a summary of a single gun-related incident on school grounds. Features which are reported on include: the date the incident occurred, the school at which the incident occurred, the city and state in which the incident occurred, whether or not the incident occurred during school hours, the number of shots fired, the type of weapons involved in the incident, details about what happened to the shooter, whether the shooter was affiliated with the school, whether the shooter had a criminal history, and the total number of victims.

Handgun and Rifle Comparison This report provides a comparison of weapon usage in gun-related incidents for a region. Features which are reported on include: the total number of all shooting incidents, the total number of incidents involving a handgun, the total number of incidents involving a rifle, and whether or not the number of incidents involving a rifle was higher than the number of incidents involving a handgun.

C.4 Environmental Sustainability

Within this domain, we generate two unique kinds of reports: Statewide Wildfire Report and County AQI and Wildfire Trends.

Statewide Wildfire Report This report provides a summary of all wildfire occurrences in a given state of the United States. Features which are reported on include: the total number of wildfires in the state each year, the average wildfire size in the state each year, the median wildfire size in the state each year, the average wildfire size in the state each year, the average wildfire duration in the state each year, the total number of wildfires in the United States each year, the average wildfire size in the United States each year, the median wildfire size in the United States each year, and the average wildfire duration in the United States each year.

County AQI and Wildfire Trends This report provides a report on air quality and wildfire occurrences for a single state of the United States. Features which are reported on include: the total number of days for which the AQI was "good" in

Document Type	Domain	Document Description
Patient Visit Report	Healthcare	Visit summary for a single stay in the emergency department
Subject Report	Healthcare	Patient summary for all emergency department visits of a single patient
MSA Housing Price Trends	Urban Housing	Housing price trends for a single metropolitan region of the U.S.
National Housing Price Trends	Urban Housing	Overall housing price trends across U.S.
Summary of Shooting Incidents	Criminal Justice	Summary of all gun-related incidents on school grounds for a single county, district, or state
Shooting Incident Report	Criminal Justice	Summary of a single gun-related incident on school grounds
Handgun and Rifle Comparison	Criminal Justice	A comparison of weapon usage in the gun-related incidents for a region
Statewide Wildfire Report	Environmental Sustainability	A summary of all wildfire occurrences in a given region
County AQI and Wildfire Trends	Environmental Sustainability	Report on air quality and wildfire occurrences for a given U.S. state
High School Expenditure / Student Performance	Education	Report on the relationship between student performance and school funding
County High School Report Card	Education	Summary of high school students' performance for a county in Illinois
Case Summary	Legal and Judicial	Summary of a single legal case
Judge Summary	Legal and Judicial	Summary of single judge's tenure
County Income Trends	Socioeconomics	Summary on income trends for a single county
County Poverty Trends	Socioeconomics	Summary on poverty trends for a single county
Business Performance Report	Business	Business performance summary for a single business
State Business Performance Report	Business	Summary of the performance of all businesses in a county

Table 5: A listing of the document types that we generate as part of our evaluation.

the state for each year, the total number of days for which the AQI was "moderate" in the state for each year, the total number of days for which the AQI was "unhealthy for sensitive groups" in the state for each year, the total number of days for which the AQI was "very unhealthy days" in the state for each year, the total number of days for which the AQI was "hazardous" in the state for each year, the maximum AQI value in the state for each year, the median AQI value in the state for each year, the total number of wildfires in the state each year, the average wildfire size in the state each year, the median wildfire size in the state each year, and the average wildfire duration in the state each year.

C.5 Education

Within this domain, we generate two unique kinds of reports: High School Expenditure and Student Performance Report and County High School Report Card.

High School Expenditure and Student Performance Report This report provides a report on the relationship between student performance and school funding. Features which are reported on include: the school name, school district, the city and county in which the school is located, the type of district the school is in, the grades served by the school, the total student enrollment, the student attendance rate, the average class size for all grades, the total federal expenditures per pupil, the total state expenditures per pupil, the total expenditures per pupil, the percentage of students demonstrating advanced performance on SAT reading, the percentage of students demonstrating proficiency on SAT reading, the percentage of students demonstrating advanced performance on SAT math, and the percentage of students demonstrating proficiency on SAT math.

County High School Report Card This report provides a summary of high school performance for a single county in Illinois. Features which are reported on include: the total number of schools in the county, the total student enrollment in the county, the median student enrollment at high schools in the county, the average student enrollment at high schools in the county, the standard deviation of student enrollment at high schools in the county, the correlation coefficient between per pupil expenditures and student enrollment, the correlation coefficient between per pupil expenditures and the percentage of students demonstrating advance performance on SAT math, the correlation coefficient between per pupil expenditures and the percentage of students demonstrating advance performance on SAT reading, and a ranking of school districts in the county according to their per pupil expenditures.

C.6 Legal and Judicial

Within this domain, we generate two unique kinds of reports: Case Summary and Judge Summary.

Case Summary This report provides a summary of a single legal case. Features which are reported on include: the name of the case, the case ID, the filing date of the case, the terminating date of the case, the year of the case, the judge presiding on the case, the city in which the case was tried, the duration of the case, the average duration for all cases, the average duration of cases for each case type, the standard deviation of case duration of all cases, and the standard deviation of case duration of cases for each case type.

Judge Summary This report provides a summary of single judge's tenure. Features which are reported on include: the name of the judge, the total number of cases this judge has presided on, the average duration of cases this judge has presided on, the number of cases this judge has presided on for each type of case, the average duration of cases this judge has presided on for each type of case, the average duration of cases this judge has presided for each year, the average duration for all cases, the average duration of cases for each case type, the standard deviation of case duration of all cases, and the standard deviation of case duration cases for each case type.

C.7 Socioeconomics

Within this domain, we generate two unique kinds of reports: County Income Trends and County Poverty Trends.

County Income Trends This report provides a summary of poverty trends for a single county in Illinois. Features which are reported on include: the population of the county for each year, the total number of unemployed people in the county for each year, the total personal income in the county for each year, the estimated median household income in the county for each year, the per capita income of the county for the most recent year, and the per capita personal income for all counties for the most recent year.

County Poverty Trends This report provides a summary of income trends for a single county in Illinois. Features which are reported on include: the population of the county for each year, the total number of unemployed people in the county for each year, the estimated number of people in poverty below the age of 17, the estimated number of people in poverty of any age, the per capita income of the county for the most recent year, and the per capita personal income for all counties for the most recent year.

C.8 Business

Within this domain, we generate two unique kinds of reports: County Business Performance Report and County Community Performance Report.

Business Performance Report This report provides a business performance summary for a single business. Features which are reported on include: the name of the business, the address of the business, the city and state the business is located in, the Yelp category for the business, the total number of reviews for this business, the average star rating for this business, the average star rating for this business for each year, and the average star rating of all businesses within the same category.

State Business Performance Report This report provides a summary of the performance of all business in a state of the United States. Features which are reported on include: the state for which businesses are being examined, the number of businesses in the state, the average rating of businesses in each category for businesses in the state, and the top ten businesses by average star rating within this state.

D Factual Accuracy by Domain

Table 6 provides a detailed breakdown of the number of factual, refuted, and confabulated claims, grouped by domain, and compared to the total number of claims made in reports in that domain.

As is evident from the table, the majority of the claims are factual and very few are refuted. It is also important to reiterate that, although the number of confabulated claims might be considered a little high, these confabulations are almost entirely represented by claims that are not necessarily incorrect, but which may be subjective or draw conclusions, often using wording like "should" or "This indicates that..."

E Examples of Generated Documents

In this section, we present examples of documents generated using our neurosymbolic method. For each example, we describe the report type, domain,

Domain	Total Claims	Factual	Refuted	Confabulated
Env. Sustainability	698	586	34	80
Healthcare	1054	983	13	56
Criminal Justice	477	332	9	137
Business	931	785	36	101
Urban Housing	360	325	10	26
Education	626	461	9	143
Socioeconomics	430	369	10	51
Legal and Judicial	596	478	7	122

Table 6: A breakdown of the total number of claims made and the number of those claims which are factual, refuted, and confabulated for each of the 8 domains we evaluated. The results are an aggregation of reports generated by in the "per plan" format for both models (GPT-4 and StableBeluga-2).

the entity instance the report focuses on, the language model used to generate the report, and the generation mode using to structure the results derived from the analytics engine prior to their inclusion in the prompt. We then show the prompt which was passed to the language model. In examples in which the prompt is too large to show easily within a single diagram, we truncate it and denote the exclusion with ellipses. We then show the report that was generated by the model.

For clarity, we have color coordinated the facts provided in the prompt with the corresponding text that was generated as part of the reports to enable quick and direct comparisons to be made between the facts we provided to the model and its outputs. It's also worth noting that these examples were chosen to highlight interesting behaviors the models exhibit when producing their generations.

The first report can be seen in Figure 4 where a report on the county poverty trends of Champaign County from 1969 to 2022 that was generated by GPT-4 is shown. Multiple times series are provided as part of the input prompt. The model has a tendency to summarize these time series with a general trend and brief summary statistics such as the minimum or maximum value or the first and last values. In general, it is preferable to have the planning module and subsequent analytic processes provide such trend values and summary statistics (since they are guaranteed to be correct based on the data), rather than rely on the model (which may make a mistake in predicting the trend or picking the min/max values) to provide summaries.

The second report can be seen in Figure 5 where a summarization of stays of the subject with ID 10023239 that was generated by using GPT-4 is shown. This report includes the time, duration and diagnosis of the six unique stays that the subject had. Note that the data had all personally identifying data removed to maintain the subject's anonymity. Consequently, most personal details are omitted and the patients are referred to by a unique ID instead of their name. Within this report, a series of diagnoses are described within the context provided to the model. These diagnoses are given in the same shorthand that was provided by the healthcare practitioner that wrote them. However, in the output the model was able to expand this shorthand into the full diagnoses. This indicates that the model has some encoding of medical knowledge that allows it to do so.

The third report can be seen in Figure 6 where a summary of a judge's tenure generated by StableBeluga-2 is shown. The report provides insights on the workload and experience of the judge. Within this report, there is an instance where the case type mentioned in the prompt is *None*. When this is discussed in the generated output, the model's outputs provides some speculation that this value was *None* "due to insufficient data." Interestingly, when the case type values are omitted elsewhere in the prompt, the model does not speculate or confabulate as to why this is the case. This raises questions as to what situations will actually cause the model to confabulate or speculate as part of its generation. This behavior warrants more thorough investigation that we leave as future work.

The fourth report can be seen in Figure 7 where a report on the AQI and wildfire trends of Carter County Missouri is shown. This report was generated by StableBeluga-2 using the *per plan* statement generation mode. Within the prompt provided to the model, there are missing values for the max AQI in the time series. This was the result of an error that occurred in the analysis that left this statement with empty spaces. Despite these values being missing, the model's generation contains a judgment that the max AQI has decreased over time, which is unsupported by the prompt. This case highlights the importance of providing the model with complete information. Without this, the model seems to confabulate summarizations and trends which may or may not actually exist.

It is worth noting that, in general, the models' generations presented the facts in the same order in which they were provided within the prompt. This property could likely be exploited to improve the structure of the reports by having a planning module determine what information is best seen first. We leave this for future work.

Report Type: County Poverty Trends
Domain: Socioeconomics
Entity Instance: Champaign County
Model Type: GPT-4
Generation Mode: Per Row

Prompt

Generate a 500 word report about county level poverty trends in Champaign County, Illinois from the given context. Use only the facts provided in the context.

Context: The year for county of Champaign County is 1969 and the resident population for county of Champaign County is None. The year for county of Champaign County is 1970 and the resident population for county of Champaign County is 163281. … The year for county of Champaign County is 2022 and the resident population for county of Champaign County is 206542. The year for county of Champaign County is 1971 and the resident population for county of Champaign County is 164300. ... The year for county of Champaign County is 1990 and the unemploed person for county of Champaign County is 2754. The year for county of Champaign County is 1991 and the unemploed person for county of Champaign County is 3372. The year for county of Champaign County is 1992 and the unemploed person for county of Champaign County is 3739. … The year for county of Champaign County is 2010 and the unemploed person for county of Champaign County is 8948. … The year for county of Champaign County is 2022 and the unemploed person for county of Champaign County is 4126. … The year for county of Champaign County is 1988 and the estimated people below 17 in poverty for county of Champaign County is None. The year for county of Champaign County is 1989 and the estimated people below 17 in poverty for county of Champaign County is 4645. … The year for county of Champaign County is 2014 and the estimated people below 17 in poverty for county of Champaign County is 9114. … The year for county of Champaign County is 2021 and the estimated people below 17 in poverty for county of Champaign County is 5292. …
The year for county of Champaign County is 1993 and the estimated people below 17 in poverty for county of Champaign County is 6338. ... The year for county of Champaign County is 1989 and the estimated people of all ages in poverty for county of Champaign County is 16911. The year for county of Champaign County is 1990 and the estimated people of all ages in poverty for county of Champaign County is None. … The year for county of Champaign County is 2011 and the estimated people of all ages in poverty for county of Champaign County is 43484. … The year for county of Champaign County is 2021 and the estimated people of all ages in poverty for county of Champaign County is 28582. …

Generated Report

This report provides an overview of poverty trends in Champaign County, Illinois, from 1969 to 2022. The data includes the resident population, the number of unemployed persons, and the estimated number of people below the age of 17 and of all ages living in poverty.

The resident population of Champaign County has seen a steady increase over the years, from 163,281 in 1970 to 206,542 in 2022. However, the data for unemployment and poverty rates only starts from 1990 onwards.

In 1990, the number of unemployed persons was 2,754, which increased to 4,126 by 2022. The highest unemployment rate was recorded in 2010 with 8,948 unemployed persons.

The estimated number of people below 17 living in poverty was first recorded in 1989 with 4,645 individuals. This number fluctuated over the years, with the highest recorded in 2014 with 9,114 individuals. By 2021, this number had decreased to 5,292.

The estimated number of people of all ages living in poverty was first recorded in 1989 with 16,911 individuals. This number saw a general increase over the years, peaking in 2011 with 43,484 individuals. By 2021, this number had decreased to 28,582.

It is important to note that the data for the resident population in 1969, the number of unemployed persons from 1969 to 1989, and the estimated number of people living in poverty from 1969 to 1988 is not available. Additionally, the data for the personal income for 2022 and the total resident population grouped by year for 2022 divided by total estimated people of all ages in poverty grouped by year for 2022 is also not available.

In conclusion, while the resident population of Champaign County has seen a steady increase over the years, the county has also experienced fluctuations in unemployment and poverty rates. The highest rates of unemployment and poverty were recorded in the early 2010s, with a gradual decrease in recent years.

Figure 4: A County Poverty Trends report generated with GPT-4.

Report Type: Subject Report
Domain: Healthcare
Entity Instance: 10023239
Model Type: GPT-4
Generation Mode: Per Plan

Prompt

Generate a 500 word medical patient summary for subject with id 10023239 from the given context. Use only the facts provided in the context.

Context: The race for id of 10023239 is WHITE. The gender for id of 10023239 is F. The count of unique stay id for subject id of 10023239 is 6. For in time grouped by stay id sorted in ascending order, the stay id for subject id of 10023239 and the in time grouped by stay id for subject id of 10023239 and the stay duration grouped by stay id for subject id of 10023239 and the acuity level grouped by stay id for subject id of 10023239 and the disease grouped by stay id for subject id of 10023239 is 30683757 and 2137-06-19 15:05:00 and 4:04:00 hours and 3 and DIAB KETOACIDOSIS IDDM, PNEUMONIA,ORGANISM UNSPECIFIED, DIAB KETOACIDOSIS IDDM, PNEUMONIA,ORGANISM UNSPECIFIED, ... , 36323598 and 2140-09-14 14:43:00 and 5:19:00 hours and 3 and Presence of insulin pump (external) (internal), Long term (current) use of insulin, Nonspec elev of levels of transamns \& lactic acid dehydrgnse, Nausea, Hyperlipidemia, unspecified, Type 1 diabetes mellitus without complications, Hypothyroidism, unspecified, Presence of insulin pump (external) (internal), Long term (current) use of insulin, Nonspec elev of levels of transamns \& lactic acid dehydrgnse, Nausea, Hyperlipidemia, unspecified,, … unspecified, 31270431 and 2140-09-17 11:35:00 and 3:25:00 hours and 3 and Other fatigue, Long term (current) use of insulin, Nonspec elev of levels of transamns & lactic acid dehydrgnse, Enlarged lymph nodes, unspecified, Type 1 diabetes mellitus without complications, Other fatigue, Long term (current) use of insulin, Nonspec elev of levels of transamns & lactic acid dehydrgnse, Enlarged lymph nodes, unspecified, Type 1 diabetes mellitus without complications, Other fatigue, Long term (current) use of insulin, Nonspec elev of levels of transamns & lactic acid dehydrgnse, Enlarged lymph nodes, unspecified, Type 1 diabetes mellitus without complications … 39348852 and 2140-10-03 06:20:00 and 2:47:56 hours and 3 and Type 1 diabetes mellitus with ketoacidosis without coma, Long term (current) use of insulin, Type 1 diabetes mellitus with ketoacidosis without coma, Long term (current) use of insulin, Type 1 diabetes mellitus with ketoacidosis without coma, Long term (current) use of insulin, Type 1 diabetes mellitus with ketoacidosis without coma, Long term (current) use of insulin, … 37578680 and 2140-10-22 18:05:00 and 22:56:00 hours and 3 and Long term (current) use of insulin, Acute kidney failure, unspecified, Type 2 diabetes mellitus without complications, Long term (current) use of insulin, Acute kidney failure, unspecified, Type 2 diabetes mellitus without complications, Long term (current) use of insulin, Acute kidney failure, unspecified, … 35942924 and 2140-11-11 11:28:00 and 5:50:00 hours and 3 and Pure hypercholesterolemia, unspecified, Long term (current) use of insulin, Type 1 diabetes mellitus without complications, Hypothyroidism, unspecified, Sarcoidosis of lung, Syncope and collapse, Pure hypercholesterolemia, unspecified, Long term (current) use of insulin, …

Generated Report

Patient ID 10023239 is a white female who has had a total of six unique hospital stays. The details of her stays are as follows:
During her first stay (Stay ID:30683757), she was admitted on June 19, 2137, at 15:05. Her stay lasted for 4 hours and 4 minutes. The acuity level during this stay was 3. She was diagnosed with Diabetic Ketoacidosis (IDDM) and Pneumonia, with the organism unspecified.

Her second stay (Stay ID: 36323598) began on September 14, 2140, at 14:43 and lasted for 5 hours and 19 minutes. The acuity level was again 3. During this stay, she was diagnosed with several conditions including the presence of an insulin pump (both external and internal), long-term use of insulin, nonspecific elevation of levels of transaminases and lactic acid dehydrogenase, nausea, unspecified hyperlipidemia, type 1 diabetes mellitus without complications, and unspecified hypothyroidism.

The third stay (Stay ID: 31270431) started on September 17, 2140, at 11:35 and lasted for 3 hours and 25 minutes. The acuity level was 3. She was diagnosed with fatigue, long-term use of insulin, nonspecific elevation of levels of transaminases and lactic acid dehydrogenase, enlarged lymph nodes, and type 1 diabetes mellitus without complications.

During her fourth stay (Stay ID: 39348852), which began on October 3, 2140, at 06:20 and lasted for 2 hours, 47 minutes, and 56 seconds, she was diagnosed with type 1 diabetes mellitus with ketoacidosis without coma and long-term use of insulin. The acuity level was 3.

Her fifth stay (Stay ID: 37578680) started on October 22, 2140, at 18:05 and lasted for 22 hours and 56 minutes. The acuity level was 3. She was diagnosed with long-term use of insulin, acute kidney failure, unspecified, and type 2 diabetes mellitus without complications.
Her sixth and final stay (Stay ID: 35942924) began on November 11, 2140, at 11:28 and lasted for 5 hours and 50 minutes. The acuity level was 3. She was diagnosed with pure hypercholesterolemia, unspecified, long-term use of insulin, type 1 diabetes mellitus without complications, unspecified hypothyroidism, sarcoidosis of the lung, and syncope and collapse.

In summary, the patient has a history of multiple hospital stays with a variety of diagnoses, including diabetes mellitus (both type 1 and 2), diabetic ketoacidosis, pneumonia, hyperlipidemia, hypothyroidism, and acute kidney failure. She also has a long-term use of insulin and an insulin pump.

Figure 5: A Subject Report generated with GPT-4.

Report Type: Judge Summary
Domain: Legal and Judicial
Entity Instance: Deborah A Robinson
Model Type: StableBeluga-2
Generation Mode: Per Plan

Prompt

Generate a nicely worded report about judge's judicial history from the given context. Use only the facts provided in the context.

Context: The name for name of deborah a robinson is deborah a robinson. The count of unique case for name of deborah a robinson is 74 cases. The average case duration for name of deborah a robinson is 647.88. The count of unique case grouped by case type for name of deborah a robinson is 2 cases, 72 cases. The average case duration grouped by case type for name of deborah a robinson is None, 647.88. The average case duration grouped by year for name of deborah a robinson is 647.88. The average case duration is 336.23. The average case duration grouped by case type is 261.79, 351.72. The standard deviation of case duration is 361.50. The standard deviation of case duration grouped by case type is 334.53, 364.96.

Generated Report

Judge Deborah A Robinson has presided over a total of 74 unique cases during her tenure on the bench. Her average case duration across all types of cases is 647.88 days. When analyzing specific case types, there are two distinct categories with an average case duration of none (likely due to insufficient data) and 647.88 days respectively. Furthermore, when examining case durations based on the year they were heard, we see that the average case duration remains consistent at 647.88 days. Breaking it down further, the average case duration for each individual case type varies between 261.79 and 351.72 days. Lastly, the standard deviation of case duration overall stands at 361.50 days, while the standard deviations for each case type category are 334.53 and 364.96 days.

Figure 6: A Judge Summary generated with StableBeluga-2.

Report Type: County AQI and Wildfire Trends
Domain: Environmental Sustainability
Entity Instance: Carter, Missouri
Model Type: StableBeluga-2
Generation Mode: Per Plan

Prompt

Generate a nicely worded report about yearly trends of air quality index and wildfire occurrence from the given context. Use only the facts provided in the context.

Context: The year for county state of Carter, Missouri and the good days grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the moderate days grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the unhealthy for sensitive groups grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the very unhealthy days grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the very unhealthy days grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the hazardous days grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the max air quality index grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the median air quality index grouped by year for county state of Carter, Missouri is . The year for county state of Carter, Missouri and the count of unique wildfire grouped by year for county state of Carter, Missouri is 1992 and 14 Wildfires, 1993 and 15 Wildfires, 1994 and 12 Wildfires, 1995 and 24 Wildfires, 1996 and 1 Wildfires, 1997 and 14 Wildfires, 1998 and 4 Wildfires, 1999 and 9 Wildfires, 2000 and 12 Wildfires, 2001 and 18 Wildfires, 2002 and 6 Wildfires, 2005 and 5 Wildfires, 2006 and 6 Wildfires, 2007 and 6 Wildfires, 2008 and 4 Wildfires, 2009 and 3 Wildfires, 2010 and 10 Wildfires, 2011 and 1 Wildfires, 2012 and 4 Wildfires, 2013 and 3 Wildfires, 2014 and 6 Wildfires, 2015 and 1 Wildfires, 2016 and 3 Wildfires, 2017 and 5 Wildfires, 2018 and 4 Wildfires, 2019 and 9 Wildfires, 2020 and 6 Wildfires. The year for county state of Carter, Missouri and the average fire size grouped by year for county state of Carter, Missouri is 1992 and 8.50, 1993 and 18.87, 1994 and 7.42, 1995 and 20.25, 1996 and 1.00, 1997 and 21.57, 1998 and 29.25, 1999 and 27.44, 2000 and 8.75, 2001 and 5.44, 2002 and 16.33, 2005 and 22.55, 2006 and 5.94, 2007 and 23.23, 2008 and 36.88, 2009 and 7.23, 2010 and 69.64, 2011 and 86.00, 2012 and 3.27, 2013 and 22.27, 2014 and 18.59, 2015 and 0.01, 2016 and 26.99, 2017 and 8.63, 2018 and 40.37, 2019 and 32.94, 2020 and 8.89. The year for county state of Carter, Missouri and the median fire size grouped by year for county state of Carter, Missouri is 1992 and 3.00, 1993 and 2.00, 1994 and 5.00, 1995 and 6.00, 1996 and 1.00, 1997 and 10.00, 1998 and 10.00, 1999 and 12.00, 2000 and 4.00, 2001 and 2.00, 2002 and 5.00, 2005 and 11.50, 2006 and 0.30, 2007 and 0.50, 2008 and 31.00, 2009 and 4.00, 2010 and 5.00, 2011 and 86.00, 2012 and 0.10, 2013 and 31.10, 2014 and 1.10, 2015 and 0.01, 2016 and 3.00, 2017 and 7.22, 2018 and 11.14, 2019 and 1.19, 2020 and 0.40. The year for county state of Carter, Missouri and the average contained date minus discovery date grouped by year for county state of Carter, Missouri is 1992 and None, 1993 and None, 1994 and None, 1995 and None, 1996 and None, 1997 and None, 1998 and None, 1999 and None, 2000 and None, 2001 and None, 2002 and None, 2005 and 0.00, 2006 and 0.00, 2007 and 0.00, 2008 and 0.00, 2009 and 0.00, 2010 and 0.10, 2011 and 0.00, 2012 and 0.00, 2013 and 0.67, 2014 and 0.17, 2015 and 0.00, 2016 and 1.67, 2017 and 0.75, 2018 and 5.33, 2019 and 0.00, 2020 and 0.00.

Generated Report

Based on the data provided, there has been an overall improvement in air quality over time in Carter County, Missouri. From 1992 to 2020, the number of "good" air quality days increased while the number of "unhealthy for sensitive groups," "very unhealthy," and "hazardous" days decreased. This indicates that efforts towards reducing pollution have had positive effects on the local environment. Additionally, the maximum Air Quality Index (AQI) value has also decreased over this period, further supporting improved air quality conditions.

Regarding wildfires, it appears that their frequency and intensity have fluctuated throughout the years. While some years experienced numerous fires with large average sizes, other years saw fewer or smaller fires. However, it's important to note that these numbers may be influenced by various factors such as weather patterns, land management practices, and human activity. Overall, though, the data suggests that wildfire occurrences are not consistently increasing or decreasing but rather show variability from one year to another.

Figure 7: A County AQI and Wildfire Trends report generated with StableBeluga-2.