Evaluating numerical factuality through question-answering in a zero-shot setting with NumerFacts as benchmark: The accuracy of large language models in numerical information recall

submitted in partial fulfillment for the degree of master of science

Mauricio Bernardo da Silva
14656779

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 31.01.2025

	UvA Supervisor
Title, Name	Jan-Christoph Kalo
Affiliation	UvA Supervisor
Email	j.c.kalo@uva.nl

ABSTRACT

This study investigates the ability of large language models (LLMs) in recalling factual numerical information in a zero-shot setting. Although there is ample research on evaluating LLMs knowledge recall, a benchmark dataset focused on numerical facts is missing. To adress this gap, we introduce the **NumerFacts** dataset, a collection of over 3,900 numerical facts across eight domains. Using a question-answering paradigm, we probed six recent open-source LLMs and analysed their accuracy, variability, and domain-specific performance.

The results show significant variability across the top performant models *Mixtral, Llama* and *Gemma*, with *Mixtral* achieving 27.14% exact matches, compared to the lowest performer *Bloom*, with only 2.44% exact matches. Compared to general factuality or commonsense knowledge benchmarks, where models achieved much higher accuracy (more than 80%), we demonstrate the substantial challenges LLMs without enhancement techniques have in recalling precise numerical facts. The **NumerFacts** dataset serves as a novel benchmark, allowing for systematic evaluation and contributing to future research in improving the factual reliability of LLMs.

KEYWORDS

large language models, numerical recalling, numbers, factuality

GITHUB REPOSITORY

https://github.com/Mau-B-Silva/NumerFacts

1 INTRODUCTION

LLMs have been transformative and influential in many domains, such as business intelligence, healthcare, legal analytics and even creative arts.[14] For example, in the study Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives, Hagos et al. describe the landscape of Generative AI and LLMs (a specific application of Generative AI). LLMs have many use cases, such as language understanding, machine translation, chatbots and text summarisation[7], and they excel in these tasks[25].

Another use case is question answering. Benchmarks such as MMLU report up to 86.4% accuracy for GPT-4 on general knowledge tasks [20]. However, factual consistency remains a challenge. The TruthfulQA benchmark finds that even top-performing models score below 30% on strict factual accuracy, while HaluEval highlights frequent hallucinations in model-generated responses[25]. Additionally, closed-book question-answering benchmarks indicate that even the most performant pre-trained models struggle with exact fact retrieval, achieving only about 21% accuracy on Natural Questions and 54% on TriviaQA[20].

When it comes to numerical capabilities, there is a clear distinction between **numerical reasoning** (the ability to perform mathematical operations and derive conclusions from them) to **numerical recalling** (the ability to retrieve accurate memorised numerical facts). As Zhao et al. highlight, these capabilities engage distinct mechanisms in LLMs: **numerical reasoning** depends on generalisable patterns learned during training, whereas **numerical recalling** requires precise retrieval of factual data, which depends on the LLM's capacity to store and recall this without error[25].

This issue of factuality is closely related to topics in the field of LLMs, namely **hallucinations**, **outdated information** and **domain-specificity**, all relevant to this research. For example, given that the outputs of these models are probabilistic in nature, hallucinations - baseless or unwarranted content - present a substantial challenge in their use as reliable knowledge sources. In certain fields, like healthcare or finance, inaccuracies in model output can lead to significant negative outcomes[20].

The factuality challenge is even more impactful to the majority of users of LLMs that do not have the resources and/or technical expertise to utilise cutting edge enhancement techniques (e.g. retrieval-augmented generation or additional fine-tuning) to improve their accuracy[5]. In some cases, even if feasible, these enhancement techniques might be impractical or costly[14].

This study builds on foundational works that examine the use of general-purpose LLMs as knowledge bases and investigate the numerical factuality of their output. To that end, a structured dataset is built from Wikidata to be referenced as a reliable collection of numerical facts, given that it is structured and well-populated[6]. The following domains were selected given the availability of numerical properties: **geography; science; history; sports; demographics; entertainment; art and literature; and personalities.** From this dataset, a structured questionnaire is derived and used to probe the ability of pre-trained LLMs (i.e. utilising the base models without further enhancement techniques, in a zero-shot approach) to accurately retrieve them.

While LLMs have demonstrated useful capabilities across diverse tasks, their capacity to accurately recall numerical information remains underexplored. This study addresses this gap by introducing the first numerical-focused dataset for evaluating LLMs, then probing and evaluating more recent LLMs with it, focusing on the following research questions:

- How accurate are LLMs in recalling factual numerical information?
- What correlates to the accuracy of numerical information recall?
- To what extent do these LLMs differ in performance when it comes to numerical information recalling?

2 RELATED WORK

The use of language models as de-facto sources of information was previously investigated by Petroni et al. in *Language Models as Knowledge Bases?*[13]. In it, the researchers assess the potential of pre-trained models, such as BERT, for factual knowledge retrieval. They introduce the LAnguage Model Analysis (LAMA) benchmark, which reports BERT-large achieving only 32.3% precision-at-one (P@1) in factual retrieval tasks. Their study also highlights that fine-tuning can substantially improve factuality, but at the risk of model overfitting. This research follows a similar approach by evaluating more recent models without additional fine-tuning to

determine if factuality has improved with scale and architecture advancements.

The exploration of prompt-based retrieval for factual knowledge extraction gained further attention in the context of pre-trained masked language models (MLMs) such as BERT and RoBERTa. However, studies have shown that MLMs' predictions often reflect dataset biases rather than true knowledge retrieval. For instance, on the uniform answer distribution WIKI-UNI dataset, the top-5 model predictions covered 52% of instances, while the actual top-5 answers only covered 7.78%, leading to significantly lower precision-at-one (P@1: 16.47 vs. 30.36 on LAMA) [3]. This suggests that factuality benchmarks may overestimate MLMs' knowledge retention. To mitigate these issues, our study adopts a question-answering paradigm (e.g., "What was the year of birth of X?") instead of template-based prompts, in an attempt to reduce reliance on biased distributions. Additionally, unlike WIKI-UNI, our dataset spans a broad numerical range, allowing for a more fine-grained evaluation of LLMs' numerical recall across varied domains and scales.

In *Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models*[24], Youssef et al. describe many methods for probing language models, all of which require a "truthful dataset" and a language model. For non-optimised inputs, the method types are cloze prompts (getting the model to predict a masked part of a statement), questions (relying on question answering datasets) and entities (inferring relational information, e.g. if two countries - the entities - share a border based on geographical coordinates). Our research will use the question answering approach described here, with a "truthful dataset" and a variety of models. Also, we will probe what the authors defined as "Vanilla PLMs", (Pre-trained Language Models - in our research, the unenhanced versions of LLMs), since this is the most straightforward approach and is said to preserve facts learned during pre-training[24]. The setting will be like a closed-book question answering, meaning that the models can only rely on information learned during pre-training, without any access to external resources (such as the internet or documents)[25]. Evaluating the models without any further enhancement is important because, as previously mentioned, these enhancement techniques are out of reach of the majority of users[5].

Kalo and Fichtel also utilise a question answering approach in *KAMEL: Knowledge Analysis with Multitoken Entities in Language Models*. Their work builds upon Petroni et al.[13] by overcoming limitations of LAMA in probing, namely the dependency on cloze-style prompts and the focus on masked language models (such as BERT, that is trained to predict missing words in a sentence). To address those challenges, the KAMEL dataset is proposed, comprised of 234 relations from Wikidata, in a question answering method, better suited for causal language models (such as GPT and other novel LLMs, trained to predict next word in a sequence but relying on previous context)[9]. Their study also proposed an evaluation procedure and evaluated many causal language models such as GPT2-XL, OPT-1.3b, OPT-6.7b,OPT-13b and GPT-J-6b, and concluded that the original LAMA benchmarks overestimate the performance of language models. The "truthful dataset" construction will follow their approach of utilising Wikidata as the source, focusing exclusively on numerical facts.

A very extensive research is provided by Wang et al. in *Survey on Factuality in Large Language Models: Knowledge Retrieval and Domain-Specificity*[20]. It highlights the crucial aspect of factuality, particularly in high-stakes domains such as health, law, and finance, where incorrect information can have significant consequences. Their findings indicate that GPT-4 achieves only 29% factual accuracy in TruthfulQA and under 50% accuracy on medical knowledge tasks[20]. This survey also extensively lists and describes evaluation metrics, grouping them into *rule-based evaluation metrics* (consistent, predictable and easy to implement, but rigid); *neural evaluation metrics* that compare model outputs to reference texts by learning evaluator models, often based on proximity of model's response to the reference; *human evaluation metrics* (better in interpretability of abstract concepts and emotional subtleties, but less scalable than automated systems and prone to subjectivity, inconsistency and error); and *LLM-based metrics* which utilise LLMs as evaluators for increased efficiency and versatility, at the cost of lack of external validation. Given these descriptions, this research will utilise *rule-based evaluation metrics* because they align best with numerical data (in a predicted/actual format) and are more scalable given the dataset structure.

Specific to numbers, the paper *Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models* by Lin et al.[11] investigates the ability of LLMs in recalling numerical commonsense. They propose the NumerSense dataset, with masked-word-prediction probes, to evaluate language models, and conclude the results are quite poor for BERT-L and RoBERTa-L models when comparing to closed-book answers from humans in 300 sampled examples. However, their research only looks at commonsense sentences with numerical information in text form, and evaluates masked language models. In contrast, this research extends that to many domains and for causal, and more novel, models, but utilise their reference of approximately 3000 truthful entries as sufficient coverage.

3 METHODOLOGY

In order to evaluate how well pre-trained LLMs perform in a zero-shot approach when prompted to recall factual numerical information, a "truthful dataset" is required. In this study, this dataset was created to contain at least 3,000 numerical facts derived from Wikidata. This dataset, named **NumerFacts**, is the basis to formulate the questions based on these numerical facts, aligning with the question-answering probing paradigm.

3.1 Building the NumerFacts Dataset

The **NumerFacts** dataset was built through a systematic process that involved the selection of domain and property, the collection of data from the query of Wikidata, the detailed cleaning and preparation and the final sampling to combine and finalise it. This section explains the step-by-step approach taken to ensure that the final dataset is balanced, robust, diverse and suitable for evaluating LLMs on numerical fact accuracy.

3.1.1 Domain and Property Selection. The first step in creating the **NumerFacts** dataset involved selecting domains and properties that would ensure diversity and coverage of numerical facts. The

availability of numerical properties (classified as 'quantity' in Wikidata, and easily filtered with the Wikidata Propbrowse tool[10]) was prioritised when defining the domains. The **eight distinct domains** and their respective *properties* are listed below:

- **Art and Literature**: *Art price, literary work publication year,* and *number of pages in books.*
- **Demographics**: *Countries' life expectancy, urban population, median income,* and *population.*
- **Entertainment**: *Film worldwide box office earnings, film budgets, number of TV series episodes, number of TV series seasons,* and *video game units sold.*
- **Geography**: *City population, country population, lake area, mountain height, mountain range length, river discharge rate,* and *river length.*
- **History**: *City populations at specific years, historical event cost of damages, duration of historical events, monuments' year of inception, number of deaths in a historical event,* and *the year a historical event started.*
- **Personalities**: *Number of children, year of birth, year of death, year they started working,* and *year they stopped working.*
- **Science**: *Astronomical object's mass, astronomical object's orbital period, astronomical object's surface gravity, element's atomic mass, element's atomic number, element's melting point, Euler characteristic of shapes, species' genome size, species' life expectancy,* and *year of scientific discovery.*
- **Sports**: *Attendance at sports event, number of athletes at sports events,* and *venue capacities.*

Some adjustments done were removing problems with value formatting (e.g., *Landmark latitudes with degrees and minutes*), or data inconsistencies (e.g., *Personalities' net worth*, where missing qualifiers such as 'M' for 'millions' led to implausible values).

3.1.2 Data Collection from Wikidata. After domain and property selection, raw data was collected using the Wikidata Query Service. Beyond the availability of properties, Wikidata was selected due to it being well populated and having good data quality when comparing to other structured knowledge graphs[6].

Each property-specific query (e.g. "elevation at sea level") returned those values for many entities (i.e. many different mountains, such as Mount Everest), under each domain (e.g. "Geography). Metadata, such as the number of sitelinks to the entity, the date when the entity was created (although all results returned empty), and last modified were also collected.

This approach resulted in one individual dataset per property (with the same value for the *Property* attribute), and the entities returned, with the respective numerical values for that property, besides other metadata and informative attributes, as mentioned above.

Challenges During Querying. One important attribute used to address challenges during querying were **sitelinks.** Sitelinks represent the number of references to an entity in Wikimedia projects[4]. It serves as a proxy for entity popularity, and was used as a filter to solve the following challenges during the data collection process:

- **Unnamed and Obscure Entities**: A non-trivial number of entries lacked meaningful labels, with the entity ID being used for the entity name (e.g., `Q21481243` as entity name) or were overly obscure (e.g. the art price of the "Histoire ancienne jusqu'à César" that had no sitelinks and no mention of who was the artist), making them unsuitable for question-answering tasks.
- **Query Timeouts**: The Wikidata Query Service enforces a 60-second limit on queries, which made the querying of many well-populated properties time out.

The process to overcome the challenges mentioned above was to progressively increase the sitelinks threshold at 0, 1, 5 or 10 sitelinks. It was done iteratively, based on the result of the queries: if the issue was still happening (e.g. query timeout at sitelinks > 1) then it would be increased (e.g. next iteration with sitelinks > 5) for the same query. Most unnamed and/or obscure entities were filtered with a > 0 or > 1 sitelink threshold, whereas > 5 and > 10 were particularly useful for preventing query timeout of well-populated properties.

Two more procedures were made to prevent query timeout, in conjunction with the sitelinks minimum threshold: specifying entity types in the query (e.g. filtering for "scientist", "politician", "artist" and "athlete" instead of any person, for the *Personalities* domain); and adding a limit to the number of entries returned query when necessary: limits used were iteratively lower from 600 to 300, in increments of 100).

Another challenge were duplicates from querying Wikidata. Many entities had multiple values listed for the same property. To prevent that as much as possible, queries were set up to first prioritise entries with Preferred Rank (representing consensus or the most reliable value)[21], then the most recent value in the absence of a Preferred Rank. However, this did not prevent duplicates completely, and further steps were taken to that end at a later step.

3.1.3 Data Cleaning and Preparation. After raw data was collected, detailed cleaning and preparation steps were performed to improve the data quality of these datasets. Each property-specific dataset was treated independently, allowing for more careful cleaning and question generation. The process was finalised with the creation of eight files, one for each domain, with each sheet as a relevant property.

Removing Duplicates. As each sheet contains a property with entities, duplicates were removed based on this entity-property pair, meaning that one particular entity (e.g. "Nile river") can appear in multiple properties (e.g. "river length" and "river discharge rate") or domains, but should be unique for each property. When an entity-property pair had multiple entries with conflicting values, due to the steps in querying failing to filter them out, all such entries were discarded instead of keeping one entry of such entity-property pair, to prevent inconsistencies and confusion during probing. For example, conflicting values for *worldwide box office earnings* of *Dead Rising* (`1,300,000` and `1,800,000`) resulted in the removal of all instances of this pair. However, if duplicate entries had multiple identical values, one was kept, as they were aligned.

Rounding Numerical Values. To standardise numerical data, a *roundedValue* column was created for each property, rounding all values to the nearest of a maximum of two decimal places. This step ensured consistency across the dataset and makes probing and analysis easier.

Removing Unnamed Entities. If entries without meaningful names (e.g., `Q21481243`) were not caught and filtered out already during collection, they were removed here. These entities are not suitable for question-answering tasks, as language models can not reasonably produce accurate responses when an entity can only be identified by their Wikidata ID.

Formatting Dates. To facilitate analysis and comparison, only **year** was extracted for dates, avoiding the issue of date formatting.

Generating Questions. For each entity-property pair, a corresponding question was generated using attributes such as *entityLabel* (the entity name), *propertyLabel* (the property label), *entityType*, *unitLabel* (provides the unit for the number), and other relevant qualifiers. All questions followed a **"what"-structured** format for clarity. The structure to generate questions was tailored for every entity-property pair so that the questions were logically coherent. Example questions include:

- **History**: *What was the population of the Amsterdam city, in 1675?*
- **Entertainment**: *What was the worldwide box office earnings of the film Good Will Hunting, in United States dollars?*
- **Demographics**: *What is the median income of the country of Luxembourg, in Euro?*
- **Geography**: *What is the elevation at sea level of the Mount Everest mountain, in metre?*
- **Personalities**: *In what year did the politician John McCain start working?*

Preventing Reasoning from Questions. In order to evaluate the recalling capabilities of LLMs, and not their numerical reasoning, entries where the question itself contained the answer were removed. For instance, historical questions such as *"In what year did the Spanish War of 1818 start?"* were excluded.

3.1.4 Weighted Sampling and Dataset Finalisation. After cleaning and preparing the data, the next step was to sample and finalise the dataset for use in the experiments. The weighted sampling process ensured balanced representation of properties within each domain and included several additional steps to refine and structure the dataset for analysis and model probing.

Weighted Sampling. The raw datasets collectively exceeded the target size of 4000 entries, but there was variation in the number of properties per domain (from 3 to 10), and the number of entities per property in these datasets (from 12 to 2913). The following action were performed to achieve a balanced dataset:

- Weighted sampling was applied to maintain the proportional representation of properties as observed in the raw data for each domain. For example, if a certain property contains 40% of all entries for its domain in the raw data, then the sampled dataset will maintain this proportion for the respective domain.
- Approximately 500 entries were sampled per domain, ensuring equal contribution from each domain.

This sampling process preserved the diversity of properties while avoiding overrepresentation of highly populated properties or underrepresentation of sparse ones.

Final Steps and Feature Engineering. Following sampling, additional steps were taken to complete the dataset:

- A *rowID* column was added to facilitate analysis during subsequent evaluations.
- The feature *days_since_last_modified* was engineered based on the *dateModified* attribute of each entry, using January 1, 2025, as the reference date. This feature will be used to assess correlations between recency and model answer accuracy.
- Finally, all empty (e.g. 'dateCreated') and irrelevant attributes (i.e. those only needed for question creation) were removed.

Dataset Outputs. Two datasets were saved at this stage:

- **NumerFacts Dataset**: This dataset represents the 'ground truth' dataset for analysis and evaluation and includes all relevant attributes, such as the entity, property, value, metadata (e.g., dateModified, sitelinks), and one engineered features (days_since_last_modified).
- **NumerFacts_Questions_Only Dataset**: This dataset contains *only* the generated questions for each entity-property pair. It is used specifically for probing the models, so as to prevent data leakage and displaying contextual information that would influence the models' output.

All of these steps were performed to ensure that the final datasets contained high quality and structured data, suitable for probing large language models.

3.2 Description Data

The **NumerFacts** dataset contains 3,929 entries, each representing a unique question based on the entity/property pair, in a total of 8 domains and 38 distinct properties (or 41, if counting by 'domain-property' pairs, since the 'population' property appears in 'Demographics', 'Geography' and 'History' domains, for different types of questions, and 'life expectancy' appears in 'Demographics' and 'Science'). The dataset contains 16 columns in total. Only the unitLabel column contains missing values, and those are for the questions where the unit was unnecessary (e.g. asking the year of inception of a monument, since the answer is just the number). The full structure is described in table 1.

Domain-Level Trends. One important insight from the domain-level analysis is the variation in *sitelinks*, which is used as a proxy for entity popularity. Figure 1 shows that most domains have their average sitelinks in the dozens. The long tailed distribution is due to the **Demographics** domain that has a significantly higher average of **313.04 sitelinks**, because the entities in this domain are countries: inherently well-linked across Wikimedia due to their relevance and cross-references, detailed in the appendix A.1.

In contrast, figure 2 displays that there is little variation for *days_since_last_modified*, ranging between **114.85** and **203.03 days** across domains, with most entities modified within the last year. It is important to note that this reflects the last modification of the entity itself, not necessarily the specific property collected for this research.

Property Analysis. Figure 3 displays the wide range between the number of entries for each property. The weighted sampling kept this variation from the raw data collection, with **NumerFacts**

Table 1: Structure of the NumerFacts Dataset

Column Name	Non-Null Count	Data Type
rowID	3929	int64
entity	3929	object
entityLabel	3929	object
domain	3929	object
property	3929	object
propertyLabel	3929	object
value	3929	float64
roundedValue	3929	float64
unitLabel	3789	object
sitelinks	3929	int64
dateModified	3929	object
entityType	3929	object
question	3929	object
fileName	3929	object
sheetName	3929	object
days_since_last_modified	3929	int64

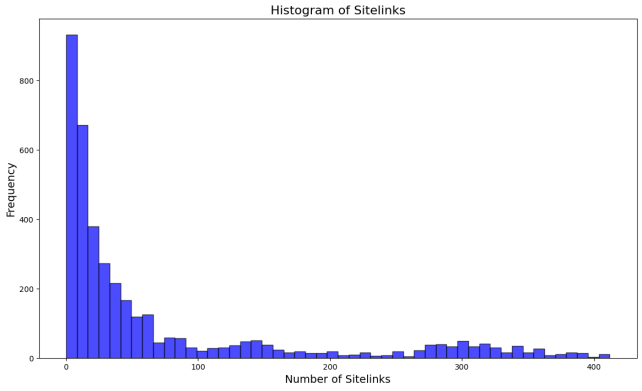


Figure 1: This histogram of sitelinks across all entries show a long tail for the number of sitelinks, with the majority of them under 100 sitelinks.

ending up with properties with number of entries as low as 1 (for 'History - event duration'), up to 373 (for 'Art & Literature - year of publication').

Values range from 0 (for 'History - event duration', that had only 1 entry, as mentioned above) to 15085750000 (for 'History - cost of damage') A.2.

Extremes in Numerical Values. **Highest Rounded Values**:

- **Hurricane Matthew** (History domain): $15.088 billion in damage costs.
- **1989 Loma Prieta earthquake** (History domain): $5.8 billion in damage costs.
- **Typhoon Soudelor** (History domain): $4.09 billion in damage costs.

Lowest Rounded Values:

- **Cave of Altacosa** (History domain): -13,000 BCE as year of inception.

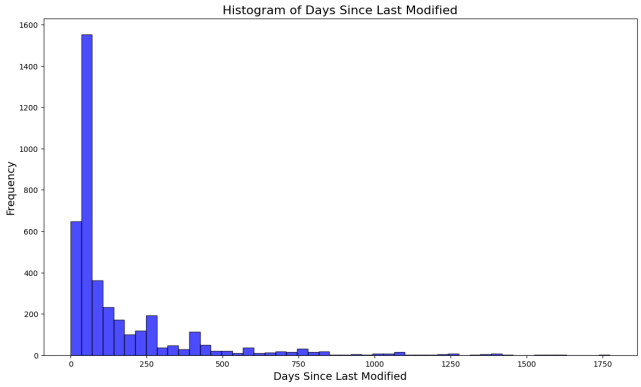


Figure 2: This histogram shows the distribution of the days since an entry was last modified. It's more heavily concentrated under 250 days, and particularly under 100 days, showcasing the recency of updates in Wikidata.

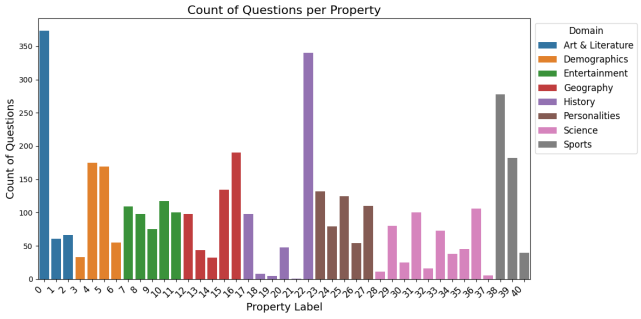


Figure 3: This bar chart shows the final number of question of each property after the weighted sampling. Properties are colour-coded based on their respective domains.

- **Bull Headed Lyre of Ur** (History domain): -2449 BCE as year of inception.
- **Henhenet** (Personalities domain): -2100 BCE as year of death.

These extreme values showcase the range of the dataset. The highest values are large monetary values, and the lowest are historical events due to the BCE convention utilised during data collection and preparation.

3.3 Experimental Setup

This section explains the experimental setup and what was the process used to probe these different LLMs with the **NumerFacts** dataset. The selected models, their configuration settings, standardisation techniques, and the automation in running the experiment are detailed below.

3.3.1 Selected Models. The following open-source models were selected for probing:

- **bigscience/bloom-7b1**[22]: A transformer-based language model created by BigScience. The name stands for BigScience Language Open-science Open-access Multilingual (BLOOM)

language model. It has 7,069,016,064 parameters on a decoder-only architecture, and it's one of the variants of BLOOM (that is much larger, with 176 billion parameters).
- **google/gemma-2-9b**[17]: This is a decoder-only large language model with 9 billion parameters developed by Google for text generation tasks, including question answering. It is part of the Gemma family, that uses the same technology as the Gemini models[18].
- **meta-llama/Meta-Llama-3-8B**[1]: The Llama 3 is a family of open-source, auto-regressive language models developed by Meta, released in April, 2024. This version has 8 billion parameters.
- **mistralai/Mixtral-8x7B-v0.1**[8]: This is a Sparse Mixture of Experts (SMoE) language model developed by Mistral AI, sharing the architecture of Mistral 7B but adding eight feed-forward blocks (experts) in each layer. It allows each token to access 47 billion parameters, but utilises only 13 billion active parameters during inference for faster processing.
- **Qwen/Qwen2.5-14B**[23][19]: This is a decoder-only, open-source model, developed by the Qwen Team, with 14 billion parameters.The Qwen Team also released other specialised models for different tasks (coding and mathematical reasoning), but the base version was chosen for a more fair comparison given the scope of the research.
- **tiiuae/falcon-7b**[2]: Developed by the Technology Innovation Institute (TII), it is a 7 billion causal decoder-only model. Its competitive performance against comparable open-source models is partly attributed to the 1.500 billion tokens of RefinedWeb data[12].

This choice was grounded on their accessibility (prioritising open-source availability), computational efficiency, prominence (high popularity in the community), performance (benchmark ratings), and recency. All selected models are hosted on HuggingFace, a data science platform and community, allowing for free access and convenient deployment, making it suitable and practical for experimentation.

3.3.2 Experimental Configuration. The following configuration was selected to improve efficiency and consistency in responses:

- **Precision and Performance Settings**: The models were loaded in 8-bit precision using the `BitsAndBytesConfig` package to optimise memory usage and computational speed, while retaining full 32-bit precision during inference to avoid potentially degrading the models performance.
- **Batch Size and Token Settings**: A batch size of 16 was set for efficient computation. The maximum token limit (`max_new_tokens`) was set to 16, which is low, but sufficient given that only concise numerical responses were asked of these models.
- **Sampling and Temperature**: Sampling was disabled, which effectively set the temperature to 0 (`do_sample=False`). This means that the outputs are more deterministic, with minimal variability.

The models were deployed and probed using Python, leveraging the HuggingFace `transformers` library for model loading and inference. Some additional packages such as `pandas` were also used for dataset manipulation, and `datasets` was utilised to more efficiently handle the **NumerFacts** questions dataset during the question answering computation.

3.3.3 Standardisation Techniques. As the models' performance will be compared against each other, standardised prompts and parsing techniques were employed, ensuring consistency:

- **Standardised Prompt**: Before each question, the following instruction was given:
  ```
  Only provide a numeric answer, no extra details.
  Do not repeat or rephrase the question. If any
  year is BC, use a negative sign (e.g., 200 BC
  → -200).
  Question: {question}
  Answer:
  ```
- **Parsing Responses**: A custom function `parse_number` was added to extract numerical values from the models' raw output. This function used regular expressions to:
  - Identify and convert years marked as BC (e.g., `200 BC` → `-200`). Also done at this stage in a different way, in case the models' output did not reflect the previous prompt instruction.
  - Extract the first relevant numerical value from the raw output, including negative numbers. This addresses the fact that LLMs can be verbose, even repeating the question or providing additional information (e.g., responses in multiple currencies when only asked about one currency).

A dedicated column (`numeric_output`) was created to store these parsed outputs, alongside the raw outputs (`raw_output`), to be later used in the analysis.

Due to memory and computational allocation constraints on the computing cluster, each model had its own (identical) script, to allow for independent execution and more efficient debugging. This approach enabled fast and standardised inference of each model with the **NumerFacts** dataset.

4 RESULTS

This section reports on the evaluation of the chosen LLMs using the **NumerFacts** dataset as the "ground truth". Before going into evaluation metrics, an assessment of response quality and appropriate preparation steps for analysis are conducted. Then, results are reported across the previously defined metrics and analysed by domain, properties, and dataset attributes.

4.1 Results Quality

Findings regarding missing values, extreme outliers, and verbosity from models are detailed here, alongside the measures taken to address them, when needed.

4.1.1 Missing Values. There was significant variability across models when it came to missing values for `extracted_number`. The main cause for missing values was identified as truncated outputs during inference, which was particularly negatively impactful for Bloom (even with the same **max tokens** setting for all models), as shown in figure 4. Other reasons were invalid responses, such as outputs without any numeric content or no answer returned at all. Qwen had no missing values, showcasing great alignment with the

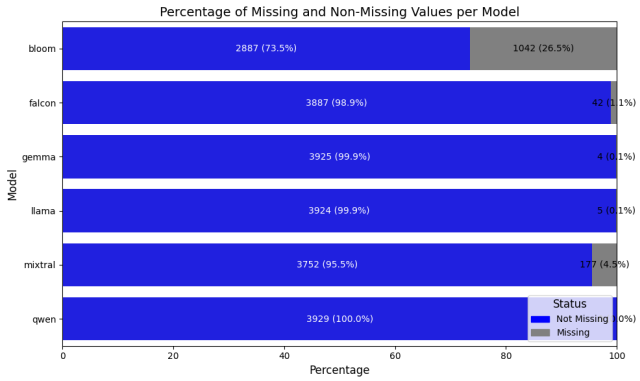


Figure 4: The absolute number of missing and non-missing numerical values after parsing displays is seen on this bar-chart, with Bloom and Mixtral being more impacted by this issue.

task. The Gemma and Llama models also performed well on this front.

To maintain the integrity of the evaluation, entries with missing numerical outputs were excluded from the calculation of evaluation metrics.

4.1.2 Extreme Outliers. Extreme outliers were classified as entries where the normalised error (already accounting for the difference in properties' scales) exceeded a threshold of 1000, meaning that the model inference was off by a factor of at least 1000 compared to the ground truth. The entries with these implausible values were removed from the dataset to prevent their disproportionate impact on evaluation metrics.

Table 2 summarises the number of extreme outliers for each model:

Table 2: Extreme Outliers by Model

Model	Extreme Outliers (Count)
Bloom	159
Falcon	64
Gemma	171
Llama	122
Mixtral	23
Qwen	55

Gemma possessed the highest number of extreme outliers (171), followed closely by Bloom (159). In contrast, Mixtral had the least amount of extreme outliers with only 23 entries, suggesting higher consistency in generating plausible predictions.

4.1.3 Alignment and Verbosity. To verify alignment with the prompt instruction to provide **only a numerical response**, verbosity was measured as the average number of extra characters in the model outputs that are not the numerical answer. Results, summarised in Table 3, reveal meaningful differences between models.

Gemma demonstrated the best alignment with an average verbosity of 17.06 characters, while Llama had the highest verbosity at 108.72 characters. On domain-level, trends indicate that verbosity was higher in Entertainment and History, with lower values in Demographics and Science.

Table 3: Average Verbosity by Model and Domain

	bloom	falcon	gemma	llama	mixtral	qwen	Domain average
Art & Literature	64.67	59.38	0.04	104.97	43.46	0.02	42.61
Demographics	65.91	49.53	2.20	70.45	38.12	0.03	37.49
Entertainment	67.97	54.94	4.19	122.19	43.60	0.11	47.42
Geography	59.62	44.30	29.25	109.69	40.49	0.03	44.52
History	58.19	42.40	32.65	128.52	41.25	0.01	50.38
Personalities	64.08	58.12	38.65	104.88	46.54	0.00	51.63
Science	65.25	46.32	21.80	107.65	37.63	0.20	46.67
Sports	69.58	54.42	5.41	116.43	47.09	0.00	48.45
Model average	64.36	51.26	17.06	108.72	42.44	0.05	46.40

4.2 Evaluation

The evaluation of LLMs in recalling numerical facts from the **NumerFacts** dataset involves comparing model outputs with ground truth values.

4.2.1 Evaluation Metrics. To comprehensively assess model performance, the following metrics are employed:

- **Percentage Exact Match (PEM)**: The proportion of model outputs that exactly match the ground truth values. For PEM, higher is better.
- **Percentage Within Tolerance (PWT)**: The proportion of model outputs that fall within a specified tolerance of the ground truth values. For PWT, higher is better. In this research, the tolerance margin (δ) was set to ±5%:

$$PWT = \frac{\text{Number of outputs satisfying } \left|\frac{y_i - \hat{y}_i}{y_i}\right| \leq \delta}{n} \times 100$$

- **Mean Relative Error (MRE)**: The average normalised difference between the model output and the ground truth, calculated as:

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

where y_i is the ground truth and $\hat{y}_i$ is the model output. For MRE, lower is better.

4.2.2 Evaluation Results. In this section, the evaluation results for the probed models are displayed, across the five defined metrics. Table 4 summarises the overall performance with micro averages: in here, the results are calculated across the whole dataset, and all entries are considered equally. Therefore, properties with more entries will disproportionately influence these calculations.

Table 4: Micro Evaluation Metrics by Model

Model	PEM (% Micro)	PWT (% Micro)	MRE (Micro)
Bloom	2.44	13.42	28.53
Falcon	11.85	28.33	8.40
Gemma	20.78	44.46	5.93
Llama	21.59	43.58	11.51
Mixtral	27.14	49.53	4.65
Qwen	17.27	35.67	3.00

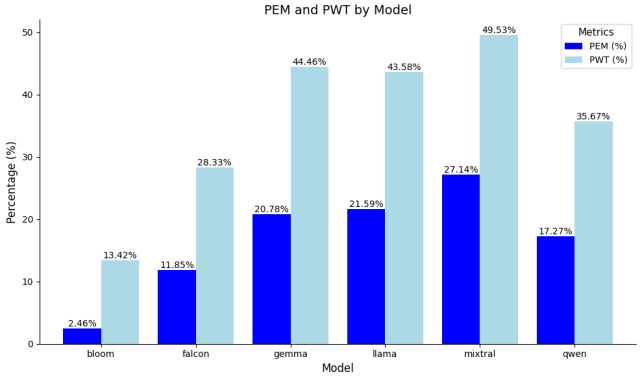


Figure 5: PEM and PWT can be seen side-by-side for each model (higher is better). Mixtral had the best results on both, followed by Gemma and Llama.

4.2.3 Percentage Exact Match (PEM). PEM evaluates the percentage of exact matches between the model outputs and the ground truth. Mixtral achieved the highest PEM (27.14%), while Bloom performed the worst (2.44%). The issue of factuality is clearly displayed here, with the best performing model only achieving about one in four answers precisely.

4.2.4 Percentage Within Tolerance (PWT). PWT measures the percentage of predictions within a ±5% tolerance of the ground truth. Mixtral again outperformed with 49.53%, followed by Gemma at 44.46%. Bloom scored the lowest at 13.42%, indicating issues in generating close approximations. For estimates within this threshold, Mixtral was able to provide one good answer out of two, approximately.

Figure 5 clearly display the comparison of PEM and PWT across the models, with Bloom having the worst performance, and Mixtral having the best, for both metrics.

4.2.5 Mean Relative Error (MRE). MRE quantifies the average normalised deviation between the predictions from the models and the ground truth. Qwen had the lowest MRE (3.00), indicating more consistent outputs, while Bloom had the highest (28.53), showing more extreme errors in responses.

4.2.6 Evaluation Across Domains. The performance with all models combined was quite different across the eight domains in both PEM and PWT metrics, as shown in Table 5. An ANOVA test was conducted at a 5% significance level, and it confirmed that these differences were statistically significant (*p-values* < .001 for both PEM and PWT).

Table 5: PEM and PWT Across Domains

Domain	PEM (%)	PWT (%)	Difference (%)
Art & Literature	39.81	76.08	36.27
Demographics	0.61	34.43	33.82
Entertainment	23.24	31.37	8.13
Geography	2.23	11.33	9.1
History	8.21	28.14	19.93
Personalities	34.62	69.74	35.12
Science	24.25	34.25	10
Sports	7.21	12.47	5.26

The domains **Art & Literature** and **Personalities** had the highest PEM and PWT values, demonstrating better model performance in retrieving numerical facts in them. In contrast, **Geography** and **Sports** had very low PEM (2.23% and 7.21%, respectively) and PWT (11.33% and 12.47%, respectively), demonstrating the greater challenge that models have in retrieving numerical facts for these domains.

Interestingly enough, the **Demographics** domain showed the lowest PEM (0.61%), although its PWT was comparatively much higher at 34.43%. This could be attributed to the fact that the entities for **Demographics** are countries, and it contains properties that can be costly and difficult to collect precisely (e.g. population of a country)[16].

ANOVA results confirm that the observed differences in PEM and PWT across domains are statistically significant (*p-value* = 0.0000 for both metrics).

4.2.7 Macro Averages. Macro averages provide a more balanced evaluation, to prevent properties with more entries from having disproportionate weight in the calculations (see above, with micro averages). This is accomplished by first computing the metrics for each domain-property pair and then averaging these values across all properties.

Table 6 shows the macro averages for PEM, PWT, and MRE across the probed models.

Table 6: Macro Evaluation Metrics by Model

Model	PEM (% Macro)	PWT (% Macro)	MRE (Macro)
Bloom	2.75	14.71	27.43
Falcon	11.55	26.64	7.77
Gemma	19.11	40.11	10.99
Llama	20.79	39.34	18.40
Mixtral	23.49	43.98	7.54
Qwen	16.13	30.67	2.83

Mixtral consistently achieves the highest macro PEM (23.49%) and PWT (43.98%), again displaying good performance. Qwen has the lowest MRE (2.83), suggesting it generates the most precise predictions on average across properties. Bloom performs poorly across all metrics, with macro PEM at just 2.75% and macro PWT at 14.71%.

4.2.8 Error Analysis. A more in-depth analysis of the errors in the predictions made by the models is shown here. Given that errors are computed as *actual minus predicted*, they were classified as either overestimations (negative error) or underestimations (positive error). Table 7 summarises the counts of overestimated and underestimated numerical responses for each model.

Bloom exhibited the largest imbalance, with significantly more underestimations than overestimations. Gemma showed the closest balance between overestimations and underestimations, reflecting a more symmetric error distribution. Qwen and Mixtral had relatively compact error distributions, indicating higher consistency in their predictions.

Bloom and Falcon exhibit a wider spread of errors (figure 6), with more extreme deviations compared to models like Mixtral and Qwen, which display tighter distributions, reflected in the aforementioned evaluation metrics.

Table 7: Percentage Overestimated and Underestimated Errors by Model

Model	Overestimated %	Underestimated %
Bloom	39.50	60.50
Falcon	44.09	55.91
Gemma	50.47	49.53
Llama	43.31	56.69
Mixtral	47.11	52.89
Qwen	39.78	60.22

Figure 6: The error distribution of each model can be seen in this violin plot. This plot has been truncated between the -1.2 to 1.2 range to make the shapes clearer (given that some values were close to -1000).

4.2.9 Correlation Analysis. A correlation analysis was performed to assess the potential relationships between prediction errors and three factors: entity popularity (measured by `Sitelinks`), recency of information (measured by `Days Since Last Modified`), and alignment with the prompt (measured by `Extra Text`). No correlation was found between these three factors and the normalised error, with the coefficient of correlation (R) being lower than 0.05 for all, in absolute terms.

Limitations:

- `Sitelinks`: This metric is a proxy for popularity, reflecting only the number of links across Wikimedia projects[4].
- `Days Since Last Modified`: The metric quantifies the days since the last change to the entity as a whole, not specifically to the queried property.
- `Extra Text`: Verbosity was already capped during model probing due to the `max_tokens` setting, limiting the range of this variable, and this not account for any other facets of model alignment, such as harmlessness or other ethical considerations[15].

4.3 Property-Level Analysis

This section provides a more detailed property-level analysis of performance trends and specific findings, based on sampled questions from selected domain-property pairs. Five domain-property pairs were chosen for each type of evaluation (PEM and PWT differences, worst, and best-performing properties) based on the

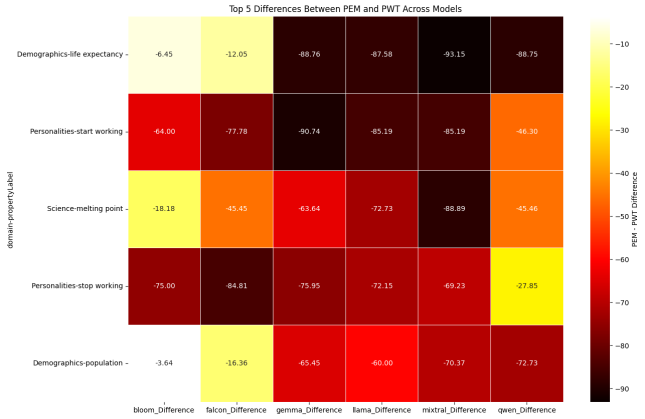


Figure 7: Darker colours signal a bigger difference between PEM and PWT. *Demographics-life expectancy* had the biggest differences across models, with the exception of Bloom and Falcon for which PWT was still low when compared to other models.

heatmaps available in the appendix A.2, with five questions per pair reviewed across all models.

4.3.1 Overall Trends. Several recurring issues and behaviors were observed throughout this analysis:

Regex and Parsing Limitations.

- **Misinterpreted "k":** Numbers with the suffix "k" (e.g., 10k) were converted to thousands (10,000). However, this caused errors when units like kilometers were included. For instance, Llama's output `"1.5 km2"` for a question about lake area was incorrectly parsed as `1500`, instead of `1.5` (ground truth: `1.83`).
- **Question repetition errors:** Some models repeated numbers from the question in their output. Falcon, for example, answered a life expectancy question with `"The life expectancy of the country of Philippines in 2018 was 70"`, leading to the capture of `2018` instead of `70` (ground truth: `71.1`). This was most common in Falcon and Bloom.

Overfitting to Prompts. Bloom and Falcon occasionally answered questions about 'years' with `200BC` `-200`. These values appeared as part of the instruction in the prompt about the expected format of the output.

Negative Outputs in Qwen. Qwen exhibited a unique issue where, occasionally, it produced a close value but incorrectly converted to negative. For example, it returned `-1475` for a historical event with a ground truth of `1473`.

4.3.2 PEM and PWT Differences. Some properties presented a significant discrepancy between PEM (exact matches) and PWT (within ±5% tolerance), as can be seen on figure 7.

Demographics – Population Estimates. For questions on population (e.g., *What is the population of Finland in 2024?*), models often produced close but not precise answers:

- Ground truth: 5,608,218
- Model responses:
  - **Mixtral:** 5,500,000 (very close)
  - **Bloom:** 2024 (incorrectly parsed the year repeated in its answer)
  - **Qwen:** 5,900,000 (close)
  - **Falcon:** 5,000,000 (close)
  - **Llama:** 5,500,000 (very close)
  - **Gemma:** 5,500,000 (very close)

Science – Melting Point of Aluminium. Questions about precise scientific constants presented some challenges due to the level of numerical detail. As an example, for the melting point of aluminium (ground truth: 660):

- **Mixtral and Qwen:** 660 (exact match)
- **Falcon, Llama, and Gemma:** Slight deviations due to rounding (e.g., 660.3).
- **Bloom:** Did not align with the question and answered with the atomic number (13) instead of the melting point.

Personalities – Career Start and Stop Years. For properties like *year start working* or *year stop working*, models often returned plausible but imprecise range of values. For example, for the year the athlete Dave Bautista stopped working (ground truth: 2019):

- **Llama:** 2019 (precise)
- Other models ranged from 2010 to 2017, which are close but not precise.

4.3.3 Worst Performing Properties. Certain properties posed significant challenges for the probed models. Some of the common challenges on them are listed below.

Overfitting to Question Structure. Models seemed to often overfit the structure of the question rather than focusing on the relevant entity. This was particularly evident in:

- **Geography – Areas of Lakes:** Despite asking about distinct lakes (*Flævatn*, *Heddersvatn*, and *Nedre Toke*), several models returned the same value:
  - **Bloom:** 1000 for all lakes
  - **Qwen:** 1.2 for all lakes
  - **Falcon:** 0 for all lakes
- **Entertainment – Number of Units Sold:** Models also provided identical responses for different videogames, such as *Sea of Stars* and *Wartales*, despite ground truth values ranging from 250,000 to 1,000,000.

4.3.4 Best Performing Properties. Some domain-property pairs had notably strong performance from many models, characterised by high precision (PEM) or consistent proximity to the ground truth (PWT).

Science – Atomic Mass. Models excelled in providing the atomic masses of elements, often producing exact or near-exact matches. For example:

- The ground truth atomic mass of Thorium is 232.04. Most models provided either the exact value or a very close approximation, such as 232.

Demographics – Life Expectancy. Life expectancy also saw strong performance, likely due to the narrow range of plausible values. For instance:

- Questions such as *"What was the life expectancy of the country of Poland in years, in 2016?"* elicited answers that were near the ground truth, with high PWT scores across models.

However, issues with **question repetition errors** persisted for Bloom and Falcon, as discussed earlier, given the structure of questions.

5 DISCUSSION

This study builds upon prior works such as Petroni et al.[13], Cao et al.[3], and Kalo and Fichtel[9], which explored the use of language models as knowledge bases. Unlike these studies, which primarily evaluated general factual knowledge using cloze-style prompts, this work focuses specifically on numerical factuality through a question-answering paradigm. Additionally, previous datasets often contained a limited scope of relations, whereas the **NumerFacts** dataset introduces a broader range of numerical facts across diverse domains, allowing for a more fine-grained analysis of LLM performance in numerical recall.

5.1 Reflection on Results and Methodology

The results demonstrate both strengths and limitations of LLMs in recalling numerical facts. Prior work suggested that LLMs struggle with numerical precision, as evidenced by the NumerSense dataset, where **BERT-Large achieved only 37.63% accuracy** and even fine-tuned models reached only **54.06%**[11]. Given these results, we anticipated that exact match rates in our study would fall below 20%. However, some models exceeded expectations, with the best-performing ones achieving **close to 1 in 4 (25%) exact matches**.

Despite these promising results, model performance varied significantly across domains. For example, models performed comparatively well in **Art & Literature** and **Personalities**. In contrast, they struggled in **Geography** and **Sports**, which may involve more dynamically changing or obscure numerical information. This aligns with prior research indicating that LLMs tend to perform better on well-documented, high-frequency knowledge but struggle with facts that require memorisation of specific numerical details[3, 11].

Further analysis revealed substantial variability across LLMs. Differences were observed not only in accuracy but also in output structure, with some models aligning more closely with prompt structure (such as Qwen) while others exhibited inconsistent formatting (such as Bloom). This was evidenced by the different proportions of missing values after parsing model outputs. Additionally, responses varied in length and verbosity, with some models providing direct numerical answers while others included unnecessary contextual explanations, making them harder to evaluate systematically.

5.2 Limitations, Dataset Reliability, and Future Work

This study provides insights into numerical factuality in LLMs, but several limitations must be noted. The **NumerFacts** dataset offers a scalable benchmark covering a wide range of numerical

facts; however, certain properties had few associated questions, limiting property-level analysis. Future work should establish a minimum question threshold or focus on well-populated properties to enhance dataset reliability. Additionally, the dataset currently lacks **time-sensitive numerical facts**, which could improve its applicability in tracking how LLMs handle ever-changing numerical information.

The question-answering paradigm used in this study differs from the cloze-style prompts traditionally employed in factual probing benchmarks such as LAMA. While QA-based evaluations provide a direct way to assess numerical recall, they can also introduce **prompt sensitivity**—where slight variations in phrasing may impact model responses. This raises concerns about reproducibility, as models may not always provide identical responses to semantically equivalent prompts. Future research should examine the impact of prompt variations and whether structured query templates impact consistency.

Several challenges were observed in model performance. **Output truncation**, particularly in Bloom, led to incomplete responses, suggesting the need for increased token limits. **Parsing inconsistencies**, such as misinterpretations of numerical abbreviations and duplicated question numbers, introduced noise, highlighting the importance of refined post-processing techniques. Additionally, reliance on a single prompt structure may have led to overfitting; future studies should introduce varied prompts to better assess model adaptability and factual consistency.

Attempts to correlate model accuracy with factors such as **entity popularity, recency, and verbosity** showed no significant relationships, indicating that architectural factors such as model size may be influential drivers of numerical factuality performance. Expanding the dataset, incorporating proprietary models, or integrating enhancement techniques such as **retrieval-augmented generation (RAG)** could offer further insights into improving numerical factual reliability.

Finally, given that numerical accuracy is often critical in domains such as **finance, medicine, and policy-making**, future work should assess the risks of LLM numerical errors in high-stakes applications. The development of structured factual retrieval techniques, alongside hybrid models that combine generative capabilities with human verification, could help mitigate risks associated with unreliable numerical responses.

6 CONCLUSION

This study introduced the **NumerFacts** dataset, the first benchmark designed specifically to probe numerical factuality in LLMs. By evaluating models across multiple domains, we observed significant variability in performance, with the best models achieving close to **1 in 4 exact matches** compared to the worst model having only about **1 in 40**.

Despite expectations of some level of numerical recall, results indicate that pre-trained LLMs struggle with precise numerical facts, suggesting that current models lack reliable representations of numerical knowledge. Correlation analysis did not present strong links between errors and dataset attributes, indicating that inaccuracies might be caused by deeper model limitations or training data.

Our findings contribute to the field by providing a structured evaluation of numerical factuality, filling a gap in existing factuality benchmarks, which often focus on general knowledge rather than numerical precision.

However, certain limitations affect the generalisability of these findings. Restricted dataset coverage and model selection suggest that broader evaluations—including proprietary LLMs and larger datasets—could bring further insights into numerical factuality. Future work should explore **scalability, output stabilisation, and prompt diversity** to further investigate LLMs' reliability as factual knowledge sources.

REFERENCES

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867 [cs.CL] https://arxiv.org/abs/2311.16867

[3] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1860–1874. https://doi.org/10.18653/v1/2021.acl-long.146

[4] Wikidata Contributors. n.d.. Help: Sitelinks. https://www.wikidata.org/wiki/Help:Sitelinks Accessed: 2025-01-05.

[5] Jennifer Dodgson, Lin Nanzheng, Julian Peh, Akira Rafhael Janson Pattirane, Alfath Daryl Alhajir, Eko Ridho Dinarto, Joseph Lim, and Syed Danyal Ahmad. 2024. Establishing Performance Baselines in Fine-Tuning, Retrieval-Augmented Generation and Soft-Prompting for Non-Specialist LLM Users. arXiv:2311.05903 [cs.IR] https://arxiv.org/abs/2311.05903

[6] Michael Faerber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2017. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web* 9 (03 2017), 1–53. https://doi.org/10.3233/SW-170275

[7] Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. arXiv:2407.14962 [cs.CL] https://arxiv.org/abs/2407.14962

[8] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG] https://arxiv.org/abs/2401.04088

[9] Jan-Christoph Kalo and Leandra Fichtel. 2022. KAMEL: Knowledge Analysis with Multitoken Entities in Language Models. In *Conference on Automated Knowledge Base Construction*.

[10] Hay Kranen. 2014. Property Browser. https://hay.toolforge.org/propbrowse/ Accessed: 2025-01-11.

[11] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds Have Four Legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-trained Language Models. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:218486812

[12] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 [cs.CL] https://arxiv.org/abs/2306.01116

[13] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases? arXiv:1909.01066 [cs.CL]

[14] Dana-Mihaela Petroșanu, Alexandru Pîrjan, and Alexandru Tăbușcă. 2023. Tracing the Influence of Large Language Models across the Most Impactful Scientific Works. *Electronics* 12, 24 (2023). https://doi.org/10.3390/electronics12244957

[15] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large Language Model Alignment: A Survey. arXiv:2309.15025 [cs.CL] https://arxiv.org/abs/2309.15025

[16] Chris Skinner. 2018. Issues and Challenges in Census Taking. *Annual Review of Statistics and Its Application* 5, Volume 5, 2018 (2018), 49–63. https://doi.org/10.1146/annurev-statistics-041715-033713

[17] Gemma Team. 2024. Gemma. (2024). https://doi.org/10.34740/KAGGLE/M/3301

[18] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] https://arxiv.org/abs/2403.08295

[19] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[20] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. arXiv:2310.07521 [cs.CL] https://arxiv.org/abs/2310.07521

[21] Wikidata contributors. 2025. Help:Ranking. https://www.wikidata.org/wiki/Help:Ranking Accessed: 2025-01-11.

[22] BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 [cs.CL] https://arxiv.org/abs/2211.05100

[23] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[24] Paul Youssef, Osman Alperen Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models. arXiv:2310.16570 [cs.CL] https://arxiv.org/abs/2310.16570

[25] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] https://arxiv.org/abs/2303.18223

Appendix A NUMERFACTS DATA DESCRIPTION

A.1 NumerFacts Dataset: Summary Statistics by Domain

Table 8: Summary Statistics of the NumerFacts Dataset by Domain

Domain	Count of Entries	Average Values	Average Sitelinks	Average Days Since Last Modified
Art & Literature	500	5,565,236.79	13.12	169.50
Demographics	432	13,161,298.62	313.04	203.03
Entertainment	499	59,426,352.59	27.47	135.86
Geography	498	1,551,434.29	48.00	156.08
History	500	53,761,288.25	89.70	148.89
Personalities	500	1,415.98	44.88	114.85
Science	500	88,651.77	57.78	189.06
Sports	500	15,815.95	9.39	115.13

The table above shows a summary of key statistics for the NumerFacts dataset across its eight domains. 8It reports, for each domain: the count of entries, average values, average sitelinks (a proxy for entity popularity), and average days since last modification (a proxy for recency). Notably, the **Demographics** domain possesses the highest number of average sitelinks, due to being comprised of well-linked entities like countries, while the **Entertainment** domain shows the highest average values due to the inclusion of properties like box office earnings.

A.2 NumerFacts Dataset: Analysis of Properties

Table 9: Detailed Summary of NumerFacts Dataset by Domain and Property

	domain	propertyLabel	Count_of_Entries	Average_Value	Minimum_Value	Maximum_Value	Value_Range	Average_Sitelinks	Average_Days_Since_Last_Modified
0	Art & Literature	latest known price	66	42149622.42409091	19.99	300000000.0	299999980.01	2.4545454545454546	248.53030303030303
1	Art & Literature	number of pages	61	305.75409836065575	4.0	1866.0	1862.0	1.7868852459016393	310.1475409836066
2	Art & Literature	year of publication	373	1942.7962466487936	659.0	2020.0	1361.0	16.863270777479894	132.50938337801608
3	Demographics	life expectancy	169	71.75603550295858	51.84	83.98	32.14	312.6686390532544	210.75147928994082
4	Demographics	median income	33	1897452.0303030303	20085.0	58008000.0	57987915.0	341.72727272727275	79.72727272727273
5	Demographics	population	55	23313230.254545454	1933.0	275439000.0	275437067.0	295.1272727272727	253.78181818181818
6	Demographics	urban population	175	24804715.98285714	5717.0	901991162.0	901985445.0	313.6114285714286	202.86857142857144
7	Entertainment	budget (capital cost)	100	36030663.8912	6000.0	250000000.0	249994000.0	39.31	152.11
8	Entertainment	number of episodes	117	48.136752136752136	1.0	862.0	861.0	14.863247863247864	126.94871794871794
9	Entertainment	number of seasons	109	2.8073394495412844	1.0	26.0	25.0	17.577981651376145	94.92660550458716
10	Entertainment	number of units sold	75	14247013.333333334	100000.0	300000000.0	299900000.0	20.506666666666668	179.34666666666666
11	Entertainment	worldwide box office earnings	98	254919914.4489796	950000.0	1027082707.0	1026132707.0	46.755102040816325	142.14285714285714
12	Geography	area	32	21.695625	0.23	250.0	249.77	6.625	183.15625
13	Geography	discharge rate	98	1062.4942857142858	0.3	41800.0	41799.7	28.581632653061224	139.0
14	Geography	elevation at sea level	44	2141.755681818182	120.0	5489.0	5369.0	13.704545454545455	125.6590909090909
15	Geography	length	134	428.62962686567164	8.4	6650.0	6641.6	28.98507462686567	138.5820895522388
16	Geography	population	190	4065040.9789473685	119.0	275439000.0	275438881.0	86.34210526315789	179.7157894736842
17	History	cost of damage	5	5336050000.0	2250000.0	15088000000.0	15085750000.0	25.6	220.8
18	History	duration	1	7.0	7.0	7.0	0.0	6.0	314.0
19	History	number of deaths	98	893.5204081632653	10.0	20000.0	19990.0	21.43877551020408	172.6734693877551
20	History	population	340	588944.4735294117	1073.0	28488200.0	28487127.0	120.80588235294118	142.43529411764706
21	History	start year	48	1455.6458333333333	-310.0	1992.0	2302.0	30.208333333333332	142.95833333333334
22	History	year of inception	8	-554.875	-13000.0	2023.0	15023.0	11.625	102.125
23	Personalities	number of children	110	1.6272727272727272	0.0	2.0	2.0	37.57272727272727	129.47272727272727
24	Personalities	start working	54	1912.7962962962963	1553.0	1995.0	442.0	52.48148148148148	155.09259259259258
25	Personalities	stop working	79	1961.1518987341772	1625.0	2022.0	397.0	40.075949367088604	103.59493670886076
26	Personalities	year of birth	125	1778.184	-68.0	1976.0	2044.0	50.664	114.048
27	Personalities	year of death	132	1722.1060606060605	-2100.0	2023.0	4123.0	45.27272727272727	93.68939393939394
28	Science	Euler characteristic	6	1.8333333333333333	0.0	9.0	9.0	45.5	153.5
29	Science	atomic mass	38	124.35552631578948	10.81	257.1	246.29000000000002	153.89473684210526	194.3684210526316
30	Science	atomic number	73	97.21917808219177	1.0	184.0	183.0	102.68493150684931	159.32876712328766
31	Science	genome size	16	2678741.4375	16000.0	32802969.0	32786969.0	6.5	399.75
32	Science	life expectancy	25	11.652000000000001	1.0	59.0	58.0	44.84	98.6
33	Science	mass	80	6.6579999999999995	0.03	165.0	164.97	19.525	211.4875
34	Science	melting point	11	1286.99	-259.14	3410.0	3669.14	186.0	32.90909090909091
35	Science	orbital period	106	1645.2586792452832	0.09	90553.02	90552.93000000001	23.71698113207547	200.33018867924528
36	Science	surface gravity	100	11808.9438	1.78	84500.0	84498.22	15.34	172.68
37	Science	year of discovery	45	1864.8222222222223	1300.0	2003.0	703.0	142.0222222222222	221.0888888888889
38	Sports	attendance	40	102845.9	6315.0	3404252.0	3397937.0	9.725	191.3
39	Sports	maximum capacity	182	20077.789615384616	49.71	120000.0	119950.29	12.631868131868131	91.42857142857143
40	Sports	number of athletes	278	503.521582733813	2.0	12986.0	12984.0	7.223021582733813	119.68705035971223

Property statistics. Table 9 summarises key statistics for the properties in the NumerFacts dataset, demonstrating the variability across domains. Interesting trends include the high median income range in the Demographics domain, averaging $1.89 million but peaking at over $58 million, and the Entertainment domain's worldwide box office earnings, which went above $1 billion for blockbuster films. The History domain has many extreme values, with the cost of damage for significant events averaging $5.33 billion, with the highest at $15 billion. Unsurprisingly, properties like atomic number in the Science domain have smaller ranges.

Property PEM and PWT heatmaps. The heatmaps on figure 8 display the PEM8a and PWT8b across all properties for each model. Values in blue are better (higher percentage). The "Personalities" domain has the highest contrast between PEM and PWT, suggesting that models do very well on *close enough* estimates, but much poorer for exact matches, and this is likely due to the comparatively smaller ranges of values for its properties. Beyond "Personalities", some other specific properties were models did well across both metrics are listed below:

- **Art & Literature:** year of publications
- **History:** start year
- **Science:** atomic mass
- **Science:** atomic number
- **Science:** year of discovery

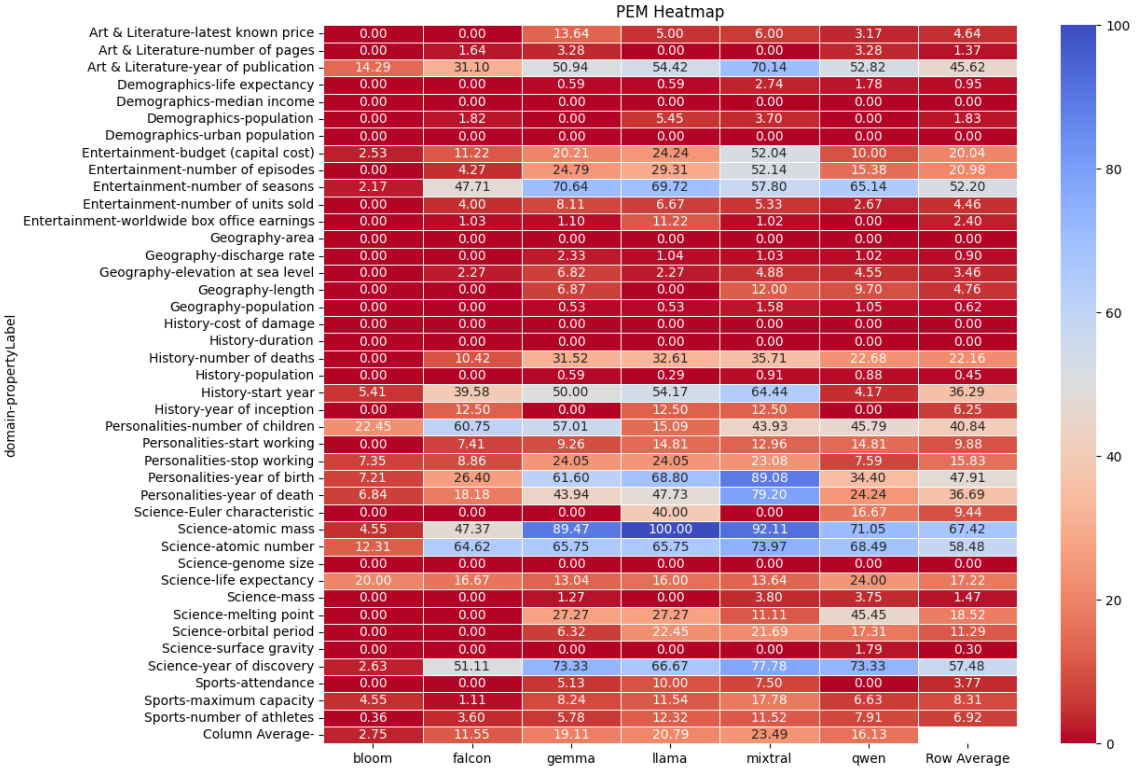
Interestingly, some properties had very poor results also for both metrics:

- **Geography:** area
- **History:** duration
- **Science:** surface gravity

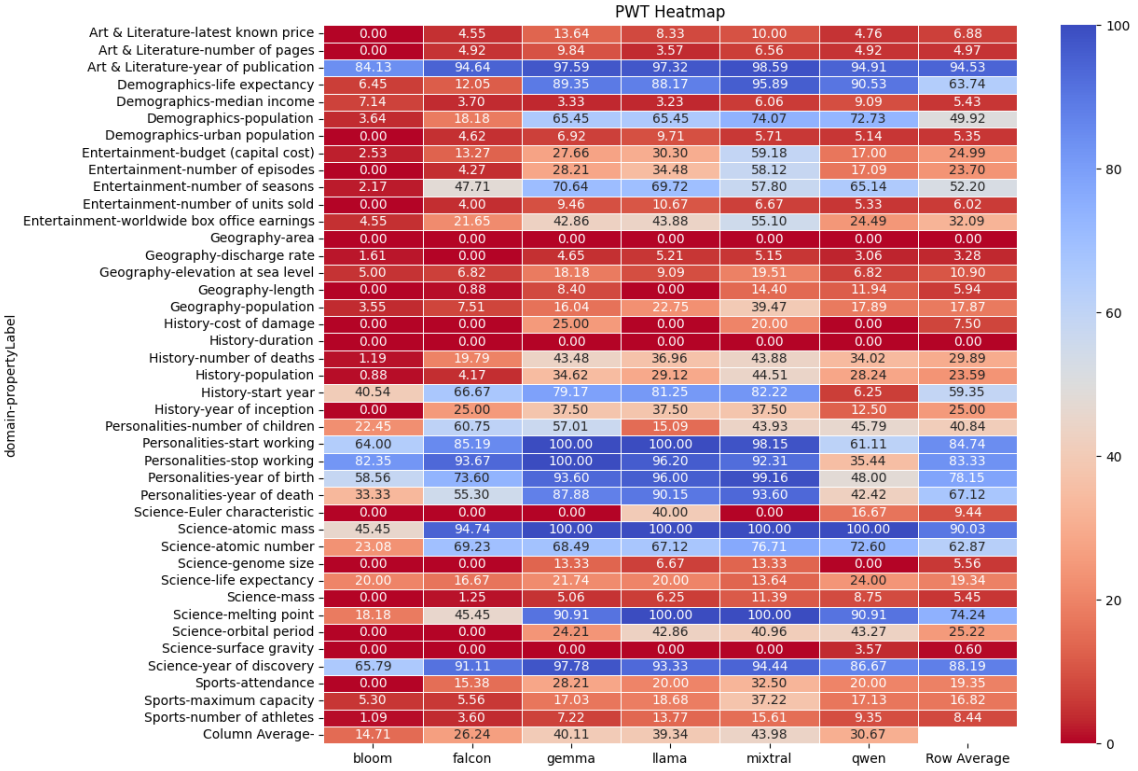
  This might hint at the differences between the type of information stored in the models, and difficulties recalling from prompts. For example, the 'duration' property for "History" had units such as seconds, minutes and years; and the 'surface gravity' property for "Science" has many astronomical objects listed, which are part of a more specific field.

Appendix B CORRELATION ANALYSIS

Scatter plots. The scatterplots in figure 9 provide a more detailed view on the normalised error plotted against sitelinks, days since last modified and extra text. The results from the correlation matrix heatmap 9d show no correlation between normalised error and any of the three factors (R values near zero).

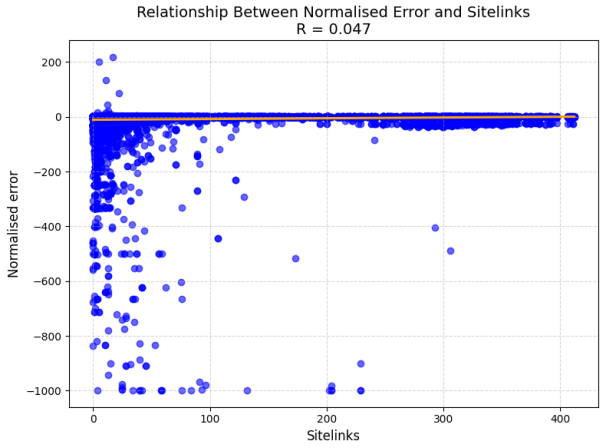


(a) This heatmap shows the PEM for all properties and models.

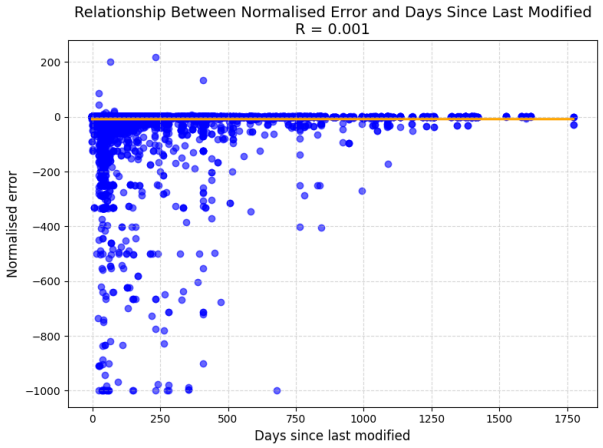


(b) This heatmap shows the PWT for all properties and models.

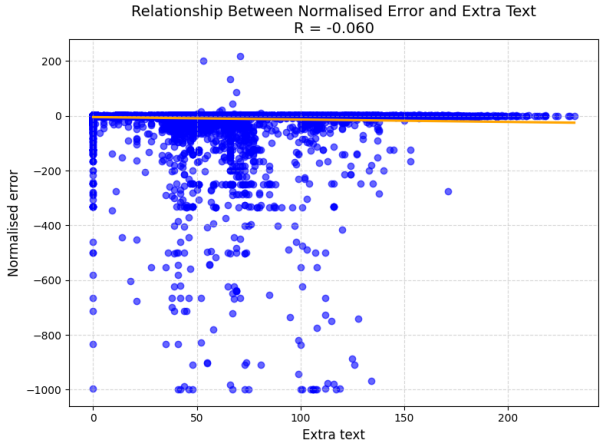
Figure 8: Heatmaps showcasing PEM and PWT metrics across all properties for each model. Blue is better.



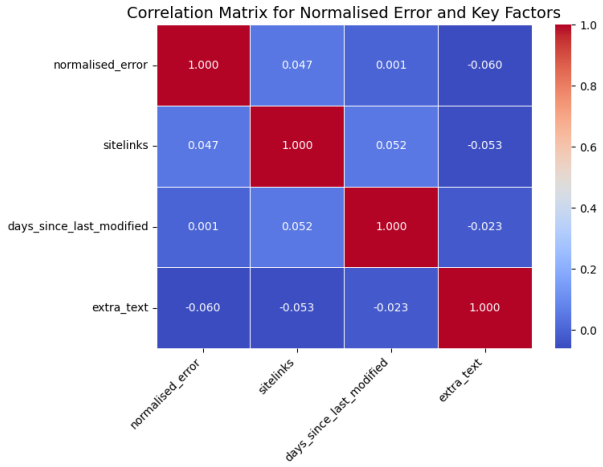
(a) This scatterplot displays the lack of correlation between normalised error and sitelinks, which is a proxy for popularity.



(b) This scatterplot displays the lack of correlation between normalised error and days since last modified, which is a proxy for recency.



(c) This scatterplot displays the lack of correlation between normalised error and extra text, a measure of verbosity and a proxy for model alignment on the task.



(d) This correlation heatmap present the correlation coefficients (R) for these paired attributes.

Figure 9: The charts above display the correlation analysis, based on scatterplots and a correlation heatmap.