The Chronicles of RIDIC: Generating Datasets with Controlled Popularity Distribution for Long-form Factuality Evaluation

Pavel Braslavski[1,2], Dmitrii Iarosh[3], Nikita Sushko[4], Andrey Sakhovskiy[4,6], Vasily Konovalov[5], Elena Tutubalina[1,5,6], Alexander Panchenko[4,5]

[1]HSE University, [2]Ural Federal University, [3]ITMO University, [4]Skoltech, [5]AIRI, [6]Sber AI

pbras@yandex.ru, {andrei.sakhovskii, a.panchenko}@skol.tech

Abstract

We present a configurable pipeline for generating multilingual sets of entities with specified characteristics, such as domain, geographical location and *popularity*, using data from Wikipedia and Wikidata. These datasets are intended for evaluating the factuality of LLMs' long-form generation, thereby complementing evaluation based on short-form QA datasets. We present the RIDIC dataset as an example of this approach. RIDIC contains 3,000 entities from three domains – rivers, natural disasters, and car models – spanning different popularity tiers. Each entity is accompanied by its geographical location, English and Chinese names (if available) and relevant English and Chinese Wikipedia content, which is used to evaluate LLMs' responses. Generations about RIDIC entities were obtained from three LLMs in English and Chinese. These were then evaluated using a third-party factuality checker, which showed that entities from our dataset caused even frontier models to hallucinate. To facilitate the evaluation of LLMs' long-form factuality in multiple languages, the code, data, and generation/evaluation scripts have been released.

Keywords: LLM evaluation, factuality, long-tail entities, multilinguality

1. Introduction

For many users, LLMs have become the go-to tool for information-seeking tasks. LLMs provide coherent answers, eliminating the need to sift through numerous documents to find, analyze, and organize information. However, LLM responses can contain factual errors, i.e., information that contradicts knowledge accumulated in reliable sources, such as Wikipedia, dictionaries, and textbooks. These errors can be especially critical in domains such as health, law, finance, and security. Factual errors are more difficult for users to detect than input- or context-conflicting hallucinations because responses look plausible, contain no obvious contradictions, and are expressed confidently (Augenstein et al., 2024).

Various approaches exist to enhance the factuality of LLMs, such as using external knowledge (RAG) and improving internal factuality with pre- and post-training techniques (Wang et al., 2025). However, in order to evaluate LLMs and track their progress, factuality evaluation benchmarks are necessary. Due to the multitasking nature and wide range of LLM applications, creating a universal benchmark is unrealistic; instead, dedicated datasets are built to evaluate different aspects of LLMs' factuality. For example, MMLU (Hendrycks et al., 2021) is designed to probe both the world knowledge and problem-solving abilities of LLMs across a wide range of tasks and domains. TruthfulQA (Lin et al., 2022) allows one to check whether LLMs have learned common misconceptions or false beliefs during training. FreshQA (Vu et al., 2024) is a dynamic benchmark comprising both evergreen and fast-changing questions designed to evaluate LLMs' up-to-date world knowledge.

LLM benchmarks also differ in format. For instance, MMLU is a multiple-choice dataset, whereas SimpleQA (Wei et al., 2024a) contains questions and short answers. One advantage of these formats is that it is relatively easy to match an LLM answer with the provided correct answer. However, it is crucial to evaluate *long-form* generation because most real-world user requests demand comprehensive and coherent responses rather than isolated facts or short answers.

At the same time, it is unclear how short-answer factuality correlates with the ability to produce longer narratives containing numerous facts (ul Islam et al., 2025). FActScore (Min et al., 2023a), a collection of person names and an eponymous evaluation framework, was seminal work on long-form factuality evaluation. Notably, the composition of the dataset is guided by the persons' popularity. As previous studies showed, LLMs' factual knowledge is strongly correlated with the popularity of pertinent entities (Kandpal et al., 2023; Mallen et al., 2023).

In this study, we present a flexible pipeline for generating datasets with the desired popularity distribution of their elements. First, we collect entities from a given class on Wikidata and select those with a Wikipedia page. Then, we calculate various popularity metrics based on Wikipedia and Wikidata. Next, we sample entities according to the chosen popularity measure and desired distribution. Finally, we collect Wikipedia data to serve as evidence for fact verification.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9893–9904
11–16 May 2026.

Dataset	Format	Code	Popularity	Size
EntityQuestions (Sciavolino et al., 2021)	QA	✗	Wikipedia links†	24k
PopQA (Min et al., 2023a)	QA	✗	Wikipedia pageviews	14k
FActScore (Min et al., 2023b)	Freeform	✗	Wikipedia pageviews/frequency	183
TriviaQA/NQ (Kandpal et al., 2023)	QA	✗	Corpus frequency†	100k+
Head-to-Tail (Sun et al., 2024)	QA	✓	Wikipedia pageviews/density	18k
WildHallucinations (Zhao et al., 2024)	Freeform	✗	Perplexity†	8k
WiTQA (Maekawa et al., 2024)	QA	✗	Wikidata triples	14k
LongFact (Wei et al., 2024b)	Freeform	✗	–	2.2k
LTGen (Huang et al., 2025)	QA	✓	Wikipedia pageviews	19k
RiDiC (this paper)	Freeform	✓	Wikipedia pageviews	3k

Table 1: Datasets with popularity facet (†*post hoc* popularity analysis). Third column indicates if the code for dataset generation is available.

As an example of our approach, we generated RIDIC – a dataset containing three types of entities: **Ri**vers, natural **Di**sasters, and **C**ar models. The dataset contains 1,000 entities of each type in three popularity tiers (head-torso-tail) based on Wikipedia pageview statistics. We gathered responses from three LLMs in two languages – English and Chinese – for these entities and evaluated their factuality. RIDIC can be seen as an extension of FActScore along three dimensions: size, domains, and languages.

Our contribution is three-fold:

1. We introduced a flexible, multilingual pipeline for generating datasets with controlled entity popularity. This enables a systematic evaluation of long-form factuality in LLMs across domains, geographies, and popularity tiers.
2. We released RIDIC, a dataset comprising 3,000 entities across three domains (rivers, natural disasters, and car models), to assess long-form factual accuracy in both English and Chinese.
3. We explored the correlation between entity popularity, location, domain, and factual precision in multilingual LLM long-form outputs. This research provides new insights into long-tail factuality challenges and multilingual evaluation reliability.

We made the code and data freely available.[1]

2. Related Work

Table 1 summarizes the main characteristics of the datasets with the popularity facet. These datasets differ in format, size, popularity proxy used, and domain. Another important difference is that some datasets have been designed to comply with a given popularity distribution of their elements, while the popularity-related analysis of others was conducted *post hoc*. For example, Kandpal et al. (2023) performed massive entity linking in a pretraining corpus, enabling them to estimate the frequency of individual entities and match them with the popular NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) datasets. Most datasets use pageview statistics from corresponding Wikipedia page as a proxy for entity popularity. Alternative approaches include link-based popularity (Head-to-Tail, EntityQuestions), the frequency of the triples containing the entity in Wikidata or another knowledge base (Head-to-Tail, WiTQA), and perplexity of the entity name based on a specific LLM (WildHallucinations).

The majority of the datasets are based on factoid questions. This format has an obvious advantage: it is relatively simple to compare the obtained answer against the reference answer. The questions come from a web search log (Natural Questions) or are generated based on Wikidata triples using templates (Entity Questions, PopQA, and Head-to-Tail) or LLMs (WiTQA and LTGen). However, the QA format differs from common practical scenarios: users typically ask LLMs more general questions and value detailed responses that reflect various aspects of a problem or entity. The FActScore dataset (Min et al., 2023a) was pioneering work that proposed an approach for evaluating the factuality of long-form LLM generation. The dataset consists of 183 person names sampled based on their popularity, as measured by Wikipedia page views and frequency in a large corpus, as well as geography. The disadvantages of this dataset are its modest size and its scope limited to biographies.

Many existing public datasets may have been contaminated by exposure to LLMs during pretraining or fine-tuning. Therefore, a recent trend is to develop an automatic pipeline for generating datasets with desired characteristics instead of creating a static dataset once (Sun et al., 2024; Maekawa et al., 2024).

Datasets for evaluating long-form generation don't contain specific questions or correct answers. They only contain a set of entities and possibly a

[1] https://github.com/s-nlp/ridic

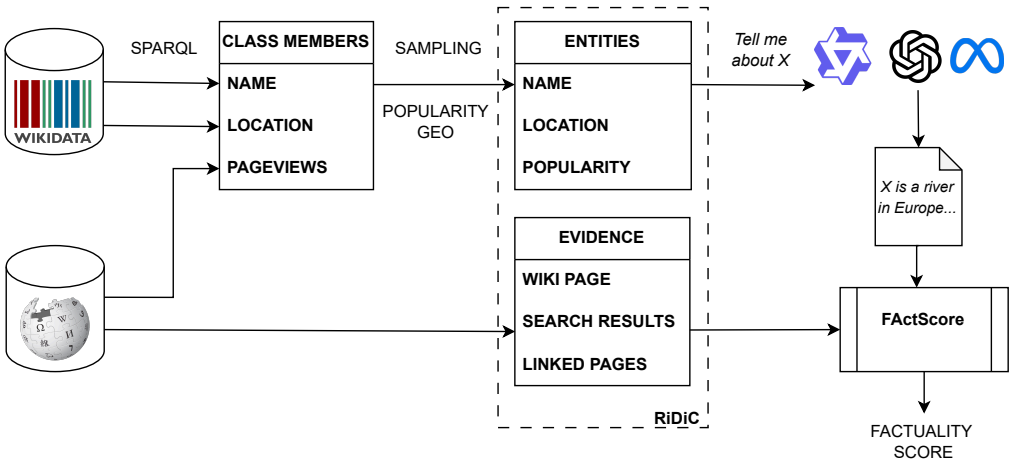

Figure 1: The process of generating a dataset begins with a Wikidata SPARQL query that defines the class of interest. Then, for each class member, attributes and popularity statistics are gathered from Wikidata and Wikipedia. The dataset is formed by sampling the required number of entities with the desired popularity distribution; Wikipedia content is collected as evidence. Once the dataset is complete, the LLMs' generated content about the collected entities is evaluated using a factuality checker.

prompt template, such as *tell me about X*. Thus, the focus shifts from generating questions to evaluating the LLM's long response. The FActScore authors conducted a manual evaluation of responses of three LLMs and proposed an automated pipeline that approximates human scores. This approach has been widely used since then. Subsequent studies have built upon this approach in various ways, including improved extraction of atomic facts, evidence retrieval using web search, and more sophisticated fact verification using multi-step reasoning (Song et al., 2024; Wei et al., 2024b; Metropolitansky and Larson, 2025).

There are several studies applying FActScore methodology to LLMs' generations about persons in non-English languages (Kim et al., 2024; Shafayat et al., 2024; Chataigner et al., 2024). These studies demonstrate that generations in higher-resourced languages are more factual. Shafayat et al. (2024) also showed that generations about people from the Western world are of higher quality, regardless of the language. These studies also indicate that the evaluation is unreliable in less-resourced languages due to lower-quality atomic fact extraction and scarcer knowledge sources that can be used as evidence.

3. Dataset Generation Pipeline

Figure 1 shows the dataset generation pipeline. The process begins with defining target classes and extracting class instances from Wikidata using SPARQL queries. Depending on the desired class, the query can be very simple (for example, the query to retrieve a list of all rivers contains just one statement `?x wdt:P31 wd:Q4022`) or more sophisticated. For convenience, the same initial query can collect the locations of the entities and the titles of their corresponding Wikipedia pages. However, when using the public Wikidata query endpoint, complex queries may fail due to timeouts. In this case, additional data can be collected for each entity via the Wikidata API.

After retrieving the basic parameters – the Wikidata identifiers, entity names (*labels*), and the titles of the corresponding Wikipedia pages in the desired languages – the popularity scores of the entities are collected. We implemented several proxies for entity popularity described in the literature: 1) a group of parameters based on Wikipedia data, such as the number of pageviews, incoming hyperlinks, edits, and page length; and 2) popularity based on the number of Wikidata triples that the entity is a part of, as either a subject or object.

Then, we divide the entities into popularity tiers and sample the desired number of entities with predefined popularity and location characteristics. In our preliminary experiments, we found that many rare entities are poorly described in Wikipedia – the corresponding page contains only one or two sentences, which hinders subsequent evaluation. Therefore, we can introduce the Wikipedia page minimal length and the Wikipedia stub flag as additional inclusion criteria.[2]

Finally, evidence is collected from Wikipedia for the sampled entities. For each entity, we add the corresponding Wikipedia page converted to

[2]A Wikipedia article is marked as a *stub* by editors if it is considered too short and incomplete.

	Rivers	Disast.	Cars
# Wikidata items	426,449	6,948	12,917
. . . w/ en label	330,789	5,890	11,300
. . . w/ en wiki page	40,914	3,706	7,347
. . . no stubs	17,783	3,413	6,445
. . . w/ zh label	15,375	1,533	2,026
. . . w/ zh wiki page	11,204	1,135	848
. . . no stubs	2,984	1,069	809
2024 pageviews	61.7M	63.4M	348.9M

Table 2: Statistics of the RIDIC classes. The framed line corresponds to the **base class** – the set upon we collect popularity statistics and define popularity tiers. Entities are sampled from a subset with more stable/complete Wikipedia pages (no stubs, next line).

	Rivers	Disasters	Cars
Head	81 (81)	20 (20)	100 (77)
Torso	200 (150)	92 (81)	200 (98)
Tail	719 (489)	888 (622)	700 (220)
Africa	217 (184)	18 (8)	0 (0)
Americas	266 (136)	246 (150)	233 (67)
AAO	264 (171)	332 (274)	381 (196)
Europe	253 (229)	103 (57)	371 (129)
Unknown	0 (0)	301 (234)	15 (3)
Total	1,000 (720)	1,000 (723)	1,000 (395)

Table 3: RIDIC dataset statistics (# of entities with Chinese Wikipedia pages in parentheses).

plain text. Previous studies have shown that using additional knowledge sources makes evaluations more reliable (Song et al., 2024; Kim et al., 2024; Wei et al., 2024b). One can expand the information included in the dataset by using pages that link to the entity's page, as well as the content of pages returned when searching for the entity's name through the Wikipedia search API. This information makes the dataset self-contained and contributes to the reproducibility of the evaluation results. A richer source of evidence (e.g., web searches) can improve coverage; however, the problem of entity disambiguation would need to be addressed in this case. Using a static knowledge base as a source of evidence may not be the best option if the entities of the selected classes change over time.

4. RIDIC Dataset

We used the proposed pipeline to generate the Rivers/Disasters/Cars (RIDIC) dataset.[3] The featured entities belong to three distinct classes: natural objects, natural events, and technical products. We collected the initial set of entities using a SPARQL query corresponding to Wikidata classes *river* (Q4022), *natural disaster* (Q8065), and *automobile model* (Q3231690).[4]

The classes differ significantly in their structure and size. For instance, there are over 400k river entities on Wikidata, but fewer than 10% of them are linked to a Wikipedia page.[5] For disasters and car models, the Wikipedia page requirement reduces the number of objects by a smaller margin, see Table 2. Despite their significant size differences, Rivers and Disasters (40.9k vs. 3.7k) attracted roughly the same amount of attention from Wikipedia users – over 60M pageviews in 2024. Cars are much more popular: 7.3k pages received nearly 349M pageviews.

We retrieved the locations of entities from Wikidata (for Cars – through their manufacturers) and aggregated them into four regions – Africa, the Americas, Asia/Australia/Oceania (AAO), and Europe. The geographic distribution of entities is also highly uneven. For example, only 88 (2.4%) natural disasters and four (0.05%) cars are attributed to Africa. Wikidata lacks the location of 1,440 (38.9%) disasters, primarily hurricanes that originate in the ocean and spread across vast territories.

We used pageviews of an entity's English Wikipedia page in 2024 as the main measure of its popularity. Based on the collected data, we divided all entities into three popularity tiers: head, torso, and tail. The cumulative number of pageviews for each tier accounts for one-third of the total number of views for the class. The three entity classes exhibit different popularity distributions. E.g., the 81 most popular rivers (0.2% of the 'base class') and the 267 most popular cars (3.6%) account for one-third of the yearly pageviews in their respective classes.

We also calculated the correlation between a battery of implemented popularity measures and found that they differ from class to class. In the Rivers, the most skewed collection with a few very popular items, we observed correlations above 0.7 in two groups of scores: 1) English Wikipedia # page edits, pageviews, and # inlinks; and 2) Chinese Wikipedia # inlinks and pageviews. The correlation between English and Chinese pageviews is 0.62/0.16/0.60 in Rivers/Disasters/Cars, respectively. In Cars, the strongest correlation is between English Wikipedia page edits and views (0.73),

[3]The name refers to *Riddick*, the protagonist of the *Chronicles of Riddick* series, who can see in the dark. We hope the dataset will improve our understanding of LLMs' capabilities in the dusk area of rare entities.

[4]Note that the *car model* class reflects different manufacturers' naming approaches. For example, each BMW 3 series version is described separately, e.g. *BMW 3 Series (E21)* (Q730915), while all six *Honda CR-V* generations are represented as a single entity (Q255461).

[5]We assume this is due to mass imports from an external database or gazetteer.

Popularity	Rivers	Disasters	Cars
Head	Rio Grande	1958 Lituya Bay earthquake and megatsunami	Xiaomi SU7
	Jordan River	Hurricane Beryl	Honda CR-V
	Yangtze	Eruption of Mount Vesuvius in 79 AD	Ferrari Testarossa
Torso	Grand River (Michigan)	2024 Spanish floods	Cadillac de Ville series
	Medjerda River	2011 Christchurch earthquake	Mini Countryman
	Solo River	Nankai megathrust earthquakes	Mitsubishi Pajero Sport
Tail	Pequonnock River	1977 Vrancea earthquake	Cadillac Brougham
	Pastaza River	2019 East Azerbaijan earthquake	Renault Avantime
	River Yare	2010 Salang avalanches	Moskvitch 402

Table 4: Example entities from the RIDIC dataset (Wikipedia titles).

while in Disasters – between English Wikipedia inlinks and pageviews (0.68). These observations suggest that different popularity proxies are not interchangeable and researchers should carefully choose from the available options.

We collected 1,000 elements from each base class. When possible, we sampled uniformly from four continents and aimed for a 100/200/700 head/torso/tail distribution. Since the heads of the Rivers and Disasters classes contain fewer than 100 entities, we included all of them and sampled additional items from the tail. To ensure that we had enough information to validate LLM responses, we filtered out entities with Wikipedia stubs and pages shorter than 200 characters. This biased the resulting collection slightly toward more popular items; otherwise, we had no data for evaluation.

In principle, all base class entities can be used for evaluation. However, we believe that 3k entities are close to the optimal size. Note that we obtain 25-30 atomic facts from LLM responses for each entity, which ensures the stability of evaluation. A larger dataset would hinder intensive experiments with different settings and models since evaluating long-form generations is a fairly resource-intensive compared to short answers.[6]

We would like to mention the issue of ambiguity that is often overlooked in other datasets with popularity facet. As we move to less popular items, we observe a higher proportion of entities with the same name. Their Wikipedia titles contain disambiguating information in parentheses, e.g. *Colorado River (Argentina)* (there are eight Colorado Rivers on English Wikipedia as of September 2025). If we omit the disambiguation information, there are 3,293/194/169 non-unique names in the base classes corresponding to Rivers/Disasters/Cars. This name ambiguity should be addressed during evaluation, especially when evidence is collected through search.

Finally, we collected information that can serve as evidence for fact verification. In addition to the Wikipedia page about the entity, we collected the top-10 results returned by the Wikipedia search API using the entity's Wikipedia title as the query and the default parameters.[7] We also collected all Wikipedia pages that link to tail entities because these rare entities typically have limited information on their main Wikipedia page, which makes factual coverage insufficient for reliable verification. Note that, in contrast to search results, linked pages provide *disambiguated* additional content in the case of namesake entities.

When generating the RIDIC dataset, we did not address the problem of the "ever-greenness" of facts about the selected entities. However, we believe that rivers, natural disasters and cars are fairly "stable" objects, the facts about which do not change much over time.

We added Chinese labels, Wikipedia titles and Wikipedia pages. The dataset can easily be expanded to include other languages. However, the main issue is the quality of factuality *evaluation* in languages other than English (Kim et al., 2024; Shafayat et al., 2024; Chataigner et al., 2024).

RIDIC statistics can be found in Table 3, examples from different classes and popularity tiers can be seen in Table 4.

5. Experiments

5.1. Factuality Evaluation

Evaluation Methodology For factuality evaluation experiments on RIDIC, we adopt a modified version of the FActScore (Min et al., 2023a) fact checking tool. FActScore implements the LLM-as-a-judge three-step pipeline: (i) *atomic fact extraction*, (ii) *evidence retrieval/ranking*, and (iii) *fact verification* against retrieved evidence. In the fact extraction step, an LLM is prompted to generate self-contained and unambiguous facts that can be verified independently of other facts or the initial text. Then, supporting evidence from an external knowledge source is retrieved and ranked. Finally, each atomic fact is verified and labeled as either 'supported' or 'not supported'. *Factual precision*

[6] In our experiments, assessing ≈3,000×30 atomic facts from one LLM on a Nvidia RTX 3090 took 36 hours.

[7] https://www.mediawiki.org/wiki/API:Search

		Rivers			Disasters			Cars		
		Llama	Qwen	GPT	Llama	Qwen	GPT	Llama	Qwen	GPT
English										
Head	avg. length	15.83	16.12	11.75	19.35	15.15	12.80	20.22	14.56	11.76
	avg. # facts	32.41	33.93	29.98	28.25	30.45	25.15	34.04	29.78	26.99
Torso	avg. length	15.27	15.75	10.64	17.91	15.65	11.82	20.40	14.80	12.27
	avg. # facts	31.55	32.80	29.15	28.89	28.36	27.51	32.41	29.31	28.60
Tail	avg. length	14.26	14.89	9.62	16.15	15.12	10.95	19.26	14.09	11.79
	avg. # facts	29.85	30.91	27.18	26.72	28.52	25.09	31.44	27.69	26.70
Total	avg. length	14.59	15.16	10.00	16.37	15.17	11.07	19.58	14.28	11.88
	avg. # facts	30.42	31.54	27.82	26.96	28.55	25.32	31.90	28.23	27.11
Chinese										
Head	avg. length	13.43	13.59	10.36	11.25	11.5	9.3	15.21	11.46	11.07
	avg. # facts	26.08	30.51	28.11	22.23	25.0	24.46	28.45	29.37	27.55
Torso	avg. length	11.94	13.56	9.74	11.75	12.56	9.24	13.38	11.63	11.52
	avg. # facts	22.60	29.67	25.35	21.91	26.81	23.45	26.27	28.95	27.43
Tail	avg. length	11.76	13.20	9.08	10.82	11.91	8.87	13.59	11.17	11.29
	avg. # facts	21.69	29.62	22.2	20.12	25.03	21.35	25.96	26.94	25.15
Total	avg. length	11.93	13.31	8.45	10.91	11.96	8.73	13.71	11.29	10.31
	avg. # facts	22.38	29.73	23.50	20.34	25.21	21.63	26.47	27.84	26.11

Table 5: Generation statistics: average response length in sentences and average number of extracted atomic facts. Note that in case of Chinese, LLMs are prompted only with entities that have a corresponding Chinese Wikipedia page, see statistics in Table 3.

LLM generation	Atomic facts	1-page	+search	+links
Although it never attained hurricane strength, Edouard brought heavy rainfall, localized flooding, and gusty winds across coastal Texas and Louisiana.	Tropical Storm Edouard brought gusty winds.	✓	✓	✓
	Tropical Storm Edouard never attained hurricane strength.	✗	✗	✓
	Tropical Storm Edouard brought localized flooding.	✗	✓	✓
	Tropical Storm Edouard brought heavy rainfall.	✗	✓	✓
Ultimately, Debby dissipated without causing severe destruction, though it brought heavy rainfall, flash flooding, and localized damage to portions of the Lesser Antilles, Puerto Rico, and the Dominican Republic, where several fatalities were recorded.	Hurricane Debby brought heavy rainfall to Puerto Rico.	✓	✓	✓
	Hurricane Debby dissipated without causing severe destruction.	✗	✗	✗
	Hurricane Debby brought heavy rainfall to the Dominican Republic.	✗	✓	✓
	Hurricane Debby brought flash flooding to portions of the Lesser Antilles	✗	✓	✓

Table 6: GPT-5 responses about two hurricanes, atomic facts extracted, and verification results.

is defined as the ratio of supported facts over the total fact count.

Implementation Details Unlike FActScore, which uses a deprecated InstructGPT (Ouyang et al., 2022) for *fact extraction* and LLama 1 (Touvron et al., 2023) for *fact verification*, we use a single model for both steps.

Specifically, we apply Llama-3.1-8B (AI@Meta, 2024) and Qwen2.5-7B (Yang et al., 2024) to English and Chinese generations, respectively.

Additionally, we re-implement LLM inference using the vLLM library (Kwon et al., 2023), resulting in an app. six times faster inference. While the original FActScore implementation retrieves supporting passages from a local Wikipedia 2023 dump, we use pages collected as a part of our dataset. Each atomic fact is concatenated with the top-5 supporting paragraphs and passed to a verification LLM, which is asked whether the given fact is supported by at least one paragraph. Specifically, we collected 1) the entity page; 2) Wikipedia search results; and 3) pages pointing to the entity page (for tail entities only). Thus, we obtain three factuality scores depending on the evidence type used (single page, +search results, +linked pages).

Factuality Evaluation in Chinese Initially, FActScore was validated by comparing its predicted factuality scores with the factuality scores on manually labeled data (Min et al., 2023a). However, the validation was performed only on English data. To assess FActScore performance in Chinese, we translated the generated facts from the FActScore dataset, as well as supporting pages, from English to Chinese using Qwen3-235B. On average, factual precision on the translated data was 7.8% lower than human judgments (53.9% vs. 61.7%). These results highlight that long-form factuality evaluation is still challenging, especially for non-English languages.

	Rivers			Disasters			Cars		
	Llama	**Qwen**	**GPT**	**Llama**	**Qwen**	**GPT**	**Llama**	**Qwen**	**GPT**
	English								
Head	0.58	0.61	0.74	0.72	0.74	0.88	0.53	0.58	0.77
+search	0.55	0.57	0.66	0.70	0.67	0.76	0.52	0.54	0.67
Torso	0.42	0.45	0.63	0.70	0.70	0.85	0.48	0.49	0.72
+search	0.43	0.45	0.60	0.65	0.61	0.72	0.48	0.46	0.64
Tail	0.25	0.27	0.50	0.43	0.42	0.63	0.32	0.35	0.63
+search	0.29	0.31	0.53	0.48	0.46	0.62	0.35	0.36	0.59
+linked pages	0.29	0.29	0.52	0.49	0.47	0.59	0.35	0.35	0.55
AAO	0.31	0.35	0.56	0.50	0.47	0.69	0.32	0.36	0.61
Africa	0.28	0.31	0.55	0.56	0.56	0.78	–	–	–
Americas	0.41	0.45	0.62	0.47	0.47	0.66	0.44	0.46	0.71
Europe	0.24	0.25	0.45	0.55	0.50	0.74	0.41	0.42	0.69
Total	0.31	0.34	0.55	0.46	0.46	0.66	0.38	0.40	0.67
+search	0.34	0.36	0.56	0.51	0.48	0.63	0.39	0.40	0.61
	Chinese								
Head	0.39	0.43	0.50	0.48	0.45	0.45	0.33	0.40	0.45
+search	0.34	0.38	0.37	0.47	0.48	0.46	0.29	0.35	0.38
Torso	0.26	0.31	0.39	0.41	0.42	0.47	0.26	0.29	0.41
+search	0.17	0.20	0.24	0.29	0.31	0.34	0.20	0.21	0.28
Tail	0.15	0.18	0.35	0.27	0.27	0.39	0.18	0.23	0.40
+search	0.09	0.10	0.15	0.12	0.10	0.15	0.15	0.16	0.27
+linked pages	0.13	0.16	0.27	0.12	0.11	0.15	0.15	0.18	0.26
Americas	0.30	0.33	0.42	0.29	0.29	0.4	0.28	0.33	0.41
Africa	0.20	0.23	0.37	0.37	0.42	0.49	–	–	–
AAO	0.23	0.29	0.42	0.33	0.33	0.47	0.23	0.28	0.42
Europe	0.16	0.18	0.34	0.47	0.42	0.53	0.32	0.34	0.44
Total	0.21	0.25	0.38	0.30	0.29	0.40	0.26	0.30	0.42
+search	0.14	0.16	0.20	0.15	0.14	0.18	0.21	0.24	0.30

Table 7: Factuality evaluation of three LLMs' generations on the RiDiC data.

5.2. LLMs' Responses

The models Qwen-2.5-7b-Instruct (Yang et al., 2024), Llama-3-8b-Instruct (AI@Meta, 2024), and GPT-5 (OpenAI, 2025) are prompted to generate responses for each of the RIDIC entities in two languages – English and Chinese. Despite differences in topic rarity (Head/Torso/Tail), the statistics are broadly similar within each language. Chinese generations are slightly shorter and contain slightly fewer atomic facts than English generations. Llama and Qwen produce comparable counts of sentences and atomic facts, whereas GPT tends to produce fewer of both, see Table 5. Excerpts from GPT-5 responses about hurricanes, extracted atomic facts, and the their evaluations using three different evidence variants can be seen in Table 6.

5.3. Factuality Evaluation Results

Table 7 shows the factuality scores for LLM generations on the RIDIC dataset, averaged by entity popularity and region. Based on these results, a few observations can be made.

LLMs Hallucinate on Rare Entities More The observed factuality scores are strongly correlated with the entity's popularity. For example, English scores preserve the Head $>$ Torso $>$ Tail ordering for all evaluated LLMs and domains. In particular, Llama shows about 2x factuality drops from 0.58/0.72/0.53 on *Head* entities to 0.25/0.43/0.32 on *Tail* for Rivers/Disasters/Cars, respectively. The same holds for Qwen generations in English. GPT-5 is slightly more robust with respect to entity popularity, with 0.74-0.88 and 0.50-0.63 score ranges on *Head* and *Tail*, respectively.

Similarly to the English evaluation, GPT-5 is consistently more factually accurate than the smaller Qwen and Llama models. The model also shows a much lower factuality gap between frequent and rare entities in Chinese. In contrast, Llama and Qwen have about 2x gap between frequent and rare entities. For instance, Llama's performance falls from 0.39/0.48/0.33 on *Head* entities to 0.15/0.18/0.35 on *Tail* for Rivers/Disasters/Cars. Overall, the results suggest that smaller LLMs struggle to memorize long-tail facts more severely due to their limited parametric capacity.

Low Factuality in Chinese In Chinese, all three models show lower factuality scores compared to English evaluation. Particularly, the total factuality scores for English range from 0.31 to 0.67 across domains, while for Chinese generations, they lie within the range of 0.21 to 0.42. Additional context from Wikipedia search seems to further mislead the evaluation by introducing lowly relevant textual passages causing a total factuality drop ranging from 0.07 to 0.18 (Llama and GPT evaluation on Rivers).

We hypothesize that this gap is caused by the following challenges. First, current LLMs tend to perform worse in non-English languages due to less intensive pre-training and fewer high-quality linguistic resources, leading to weaker generation capabilities and lower factual precision in Chinese (Li et al., 2024).

Second, the gap can be attributed to the limited reference data available on Chinese Wikipedia, which reduces the reliability and scope of evidence for fact verification (He et al., 2025). Overall, the findings indicate current challenges in generating and verifying accurate long-form output in non-English languages and reveal the need for improved multilingual LLM capabilities and factuality evaluation methods.

Factuality is Sensitive to Domain Factuality scores vary notably across the three domains. *Disasters* and *Cars* receive the highest factuality scores from all LLMs, and *Rivers* receive the lowest factuality scores. Therefore, an LLM's ability to generate factually accurate texts depends on the target domain, but the domain complexity is generally consistent across different LLMs.

LLMs Exhibit no Clear Geographic Bias The effect of an entity's location on the factual accuracy is mixed. English responses about American Rivers and Cars are more accurate, but responses about Disasters show a different picture (note that RIDIC contains no cars and very few disasters attributed to Africa, so due to the sample size, the final scores may be skewed). It is possible that location's impact is more significant at the level of individual countries, as shown by (Shafayat et al., 2024), rather than across entire continents.

Smaller LLMs are Less Accurate GPT-5 achieves the highest factuality scores across all domains and popularity levels, indicating superior factual precision in long-form generation. The Llama and Qwen models exhibit a similar level of factual hallucinations, which suggests that models of comparable size have similar capacities for memorizing factual knowledge.

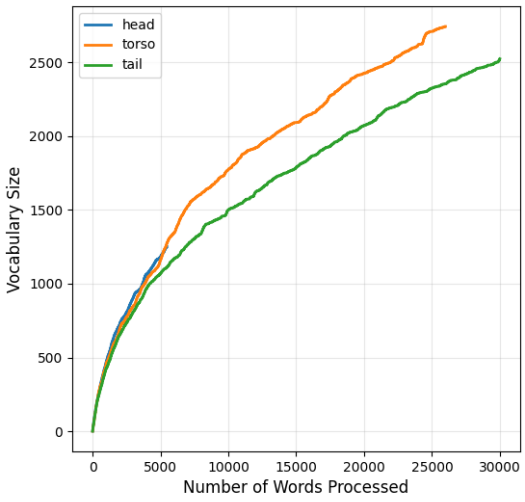

Figure 2: Vocabulary growth in Qwen's generations about disasters in different popularity tiers.

Richer Evidence's Harm and Benefit For English, incorporating Wikipedia search results and linked pages enhances factual precision for *Qwen* and *Llama* models on *Tail* entities. However, this additional context reduces performance for *Head* and *Torso* entities, as their primary pages already appear to provide sufficient context for factuality assessment. In contrast, Chinese performance drops across all frequency groups when supplementary pages are added, despite the generally less comprehensive nature of Chinese Wikipedia content. This observation may be caused by poorer evidence ranking in Chinese, resulting in misleading contexts being retrieved at high ranks.

Model	**Tier**	**#**	**RMSE** (km)	**MAPE**
Llama	Head	81	84	6.62
	Torso	187	216	10.51
	Tail	496	223	63.38
Qwen	Head	78	86	2.53
	Torso	187	143	21.46
	Tail	507	256	80.17
GPT5	Head	76	51	10.11
	Torso	179	67	3.76
	Tail	420	99	16.90

Table 8: Errors in river lengths: English generations vs. Wikidata. Note that the *relative* error (MAPE) is smaller in case of popular rivers – they are in general longer.

Targeted Evaluation In addition to the FActScore-based evaluation, we conducted an evaluation focused on a single attribute – river length. To this end, we extracted river lengths from Wikidata (property *P2043*), detected the lengths in the LLM responses using a regular expression,

and compared them. In some cases, we observed discrepancies between Wikidata and Wikipedia. For example, as of September 2025, Wikipedia indicated the length of the Chicago River as 156 miles, whereas it was erroneously indicated as 1.6 miles in Wikidata. This observation again highlights the dependence of factuality evaluation on the knowledge source used.

The results are presented in Table 8. This approach demonstrates that at least some aspects of long-form generations allow for more fine-grained evaluations than binary decisions on atomic facts.

Vocabulary Richness We investigated another aspect of LLM generations: their lexical diversity. To this end, we measured vocabulary growth (the number of unique words) as a function of text length (concatenated LLM responses) across classes and popularity tiers – the relationship described by the Heaps' law (Heaps, 1978). Our results suggest that generations about less popular entities have lower lexical diversity. This effect is more pronounced in the case of Disasters (see Figure 2) and Rivers, but is almost nonexistent in Cars. Among the three LLMs, GPT exhibits the highest lexical diversity, followed by Qwen and Llama.

6. Conclusion

This paper introduces a highly configurable, multilingual pipeline designed to generate datasets with controlled entity popularity distributions. This pipeline is intended to facilitate the long-form factuality evaluation of LLMs. Our approach allows for the systematic sampling of entities from various domains, geographic regions, and popularity tiers. We also introduce RIDIC, a dataset comprising 3,000 entities from three domains.

Our experimental results demonstrate that current LLMs exhibit significant variability in factual precision based on entity popularity, language, and domain, demonstrating notable weaknesses with long-tail entities and in non-English scenarios. Thus, the RIDIC dataset and tools provide valuable resources for advancing automatic factuality assessment and accelerating progress in trustworthy language generation. The study also revealed challenges in evaluating non-English generations. Future research would benefit from extending these methods to a more diverse set of domains, languages, and LLMs.

Limitations

While the RIDIC dataset and generation pipeline offer a robust framework for evaluating long-form factuality in large language models, several limitations must be acknowledged. First, experiments are conducted on only three LLMs (two of which are of a similar size) and in only two languages: English and Chinese.

Second, factuality evaluations for non-English languages are less reliable due to the scarcity of high-quality reference data, as well as the lower performance of current LLMs in these languages.

Third, excluding Wikipedia stubs and short articles results in a dataset that is modestly biased towards popular entities, thereby underrepresenting the long tail of entity popularity.

Furthermore, as English Wikipedia pageviews are used as a popularity signal, the popularity distribution may be biased. Finally, reliance on Wikipedia as a primary source may limit the range of evidence covered. Future work should address these issues by improving multilingual evidence collection, refining approaches to resolving ambiguity, and extending the pipeline to incorporate more diverse and up-to-date knowledge sources.

Acknowledgments

Pavel Braslavski acknowledges the support he received from the Basic Research Program of HSE University for conducting hallucination detection experiments and preparing the manuscript. The experiments were partially run on HSE computational facilities (Kostenetskiy et al., 2021).

The work of Alexander Panchenko on developing methodology of this work was supported by the Russian Scientific Foundation project № 25-71-30008 "Laboratory for reliable, adaptive, and trustworthy Artificial Intelligence".

7. Bibliographical References

AI@Meta. 2024. Llama 3 model card. *Meta*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram A. Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu,

Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Dekai Sun, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. Chinese simpleqa: A chinese factuality evaluation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19182–19208. Association for Computational Linguistics.

Harold Stanley Heaps. 1978. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wenyu Huang, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, and Jeff Z. Pan. 2025. Prompting large language models with knowledge graphs for question answering involving long-tail facts. *Knowledge-Based Systems*, 324:113648.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Vu Trong Kim, Michael Krumdick, Varshini Reddy, Franck Dernoncourt, and Viet Dac Lai. 2024. An analysis of multilingual FActScore. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4309–4333, Miami, Florida, USA. Association for Computational Linguistics.

Pavel Kostenetskiy, Roman Chulkevich, and Viachesvlav Kozyrev. 2021. Hpc resources of the higher school of economics. *Journal of Physics: Conference Series*, 1740(1):012050.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *CoRR*, abs/2404.11553.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. 2024. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5506–5521, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 6996–7045. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023a. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Min, Sewon and Krishna, Kalpesh and Lyu, Xinxi and Lewis, Mike and Yih, Wen-tau and Koh, Pang and Iyyer, Mohit and Zettlemoyer, Luke and Hajishirzi, Hannaneh. 2023b. *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. Association for Computational Linguistics.

OpenAI. 2025. GPT-5 system card.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Haeyun Oh. 2024. Multi-fact: Assessing factuality of multilingual llms using factscore. In *Conference on Language Modeling (COLM 2024)*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavaš. 2025. The curious case of factual (mis)alignment between llms' short- and long-form answers.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Qipeng Guo, Xiangkun Hu, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Xuming Hu, Zehan Qi, Wenyang Gao, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2025. Survey on factuality in large language models. *ACM Computing Surveys*, 58(1).

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *CoRR*, abs/2411.04368.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su,

Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Raghavi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *CoRR*, abs/2407.17468.