On Generalization across Measurement Systems: LLMs Entail More Test-Time Compute for Underrepresented Cultures

Minh Duc Bui[▽]   **Kyung Eun Park**[♣]   **Goran Glavaš**[◇]
Fabian David Schmidt[◇]   **Katharina von der Wense**[▽♠]

[▽]Johannes Gutenberg University Mainz, Germany   [♣]University of Mannheim, Germany
[◇]Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
[♠]University of Colorado Boulder, USA

{minhducbui, k.vonderwense}@uni-mainz.de   kyung.eun.park@uni-mannheim.de
{fabian.schmidt, goran.glavas}@uni-wuerzburg.de

Abstract

Measurement systems (e.g., currencies) differ across cultures, but the conversions between them are well defined so that humans can state facts using any measurement system of their choice. Being available to users from diverse cultural backgrounds, large language models (LLMs) should also be able to provide accurate information irrespective of the measurement system at hand. Using newly compiled datasets we test if this is the case for seven open-source LLMs, addressing three key research questions: (RQ1) What is the default system used by LLMs for each type of measurement? (RQ2) Do LLMs' answers and their accuracy vary across different measurement systems? (RQ3) Can LLMs mitigate potential challenges w.r.t. underrepresented systems via reasoning? Our findings show that LLMs default to the measurement system predominantly used in the data. Additionally, we observe considerable instability and variance in performance across different measurement systems. While this instability can in part be mitigated by employing reasoning methods such as chain-of-thought (CoT), this implies longer responses and thereby significantly increases test-time compute (and inference costs), marginalizing users from cultural backgrounds that use underrepresented measurement systems.

1 Introduction

The user base of large language models (LLMs) expands to people from a wide range of cultural backgrounds, with each person bringing distinct perspectives to their interactions with these ubiquitous systems. Intuitively, users expect that data can be *correctly* presented in ways that align with their cultural context, including measurement systems that they are used to. For instance, individuals accustomed to the metric system might find it challenging to interpret imperial units.

Prior work shows that LLMs encode factual information, including plethora of measurement val-
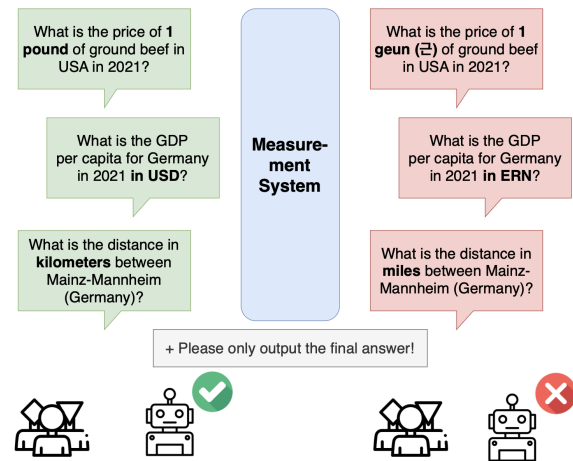

Figure 1: We test whether LLMs can translate facts across different measurement systems and show that LLMs often struggle to provide accurate information for cultural contexts underrepresented in training data.

ues, within their parameters (Petroni et al., 2019; Geva et al., 2023; Allen-Zhu and Li, 2024, *inter alia*). However, while an LLM may accurately present facts within one cultural context (presumably the one that dominated their respective training corpora), an inclusive system should be able to do so in *all* contexts. By failing to do so, LLMs would disadvantage users from cultural backgrounds that are less represented in their training data. This raises a critical, yet underinvestigated question: can LLMs effectively generalize across socio-cultural contexts, offering same levels of factual correctness irrespective of the concrete culturally-bound factors such as measurement systems? Or do they exhibit biases that hinder socio-cultural adaptability, preventing them from equitably providing accurate information to all users?

In this work, we demonstrate that LLMs are less reliable when the users' measurement system differs from the dominant system of the LLM. We find that LLMs typically default to the measurement that fits the cultural context of the data, e.g., reporting fiscal values in USD as well as length and

21262

weight in the metric or imperial system, depending on the context of the data. For instance LLMs report food prices of US goods in USD normalized per pounds. The main issue arises when users naively prompt the models with systems that differ from the dominant one, as LLMs often fail to successfully translate values from one measurement into another (e.g., monetary amounts from USD to SDG). While advanced prompting techniques such as chain-of-thought reasoning largely mitigate these conversion errors (e.g., from 72% to almost no difference for currency for the largest LLMs in our evaluation), we need to note that this improvement comes at a substantial cost. Under CoT, models produce much longer, more verbose responses as they triangulate the conversion process step-by-step. For example, obtaining the correct value for the GDP of a country in Eritrean Nakfa can increase the inference cost by up to 300%[1] compared to requesting the same value in US Dollars. Such issues have severe implications for users from marginalized cultural backgrounds, who are meant to benefit the most from equitable access to language technology. When LLMs misrepresent or inaccurately convert critical measurements (e.g., wrongly measuring monetary value of consumer goods), they alienate users by failing to align with their cultural context. Moreover, as these users frequently come from economically disadvantaged regions, they are disproportionately affected by the increased inference costs associated with extended model responses.

We address the following research questions in this work (together with respective contributions).

(RQ1) *What measurement do LLMs default to?*

We curate datasets that span diverse cultural contexts, including fiscal data, food prices, and city distances across different countries and regions. We find that LLMs default to the measurement that fits the cultural context of the data. When queried on distances between German and US cities, they provide estimates in the metric and imperial units, respectively. Likewise, LLMs express fiscal data in USD, the common convention for financial data.

(RQ2) *Do LLMs' answers and their accuracy vary across measurement systems?*

We test whether varying measurement systems affects factual accuracy of answers provided by LLMs (as illustrated in Figure 1). We find that

LLMs exhibit significant variability in behavior depending on the measurement system embedded in the prompt. **Models perform best when the query aligns with the measurement system that fits the cultural context of the data (e.g., provided by a location or entity mentions)**. For instance, models show a strong decline in accuracy when asked about food prices in pounds for a country that uses kilograms.

(RQ3) *Can LLMs mitigate these challenges with reasoning?*

We find that LLMs often struggle to generate accurate facts across different measurement systems. Nevertheless, very large LLMs (70B parameters) can successfully perform sequential single-hop inferences, starting with the default measurement and then converting the value to the target unit. This exemplifies that LLMs necessitate additional guidance to reliably triangulate to the correct value in the target measurement. Explicitly prompting LLMs to reason across measurement systems similarly stabilizes performance. Yet, such reasoning remains anchored in the default measurement, increasing inference costs for queries reflecting non-default cultural contexts.

2 Measurement Systems & Datasets

2.1 Measurement Systems

Each dataset is rooted in a specific cultural context. For instance, maps often use local units of measurement, international fiscal data such as GDP is typically reported in USD, and food prices are commonly expressed in local currencies. While previous studies have explored the geographical biases in LLMs' factual retrieval (Manvi et al., 2024; Moayeri et al., 2024), there has been limited investigation into whether these models can effectively adapt facts to diverse socio-cultural contexts. For example, can a model adapt fiscal data to a different currency or report city distances using an alternative length system? To address this gap, we focus on three measurement systems: **currency, length systems, and weight systems**. In Appendix A.1, we describe each system in depth and report how often each remains in use nowadays.

2.2 Datasets

Table 1 provides an overview of the measurement systems employed in each dataset, including details on their total sizes and representative examples. This dataset is released under the

---

	Fiscal Data	City Distances	Food Prices per Weight
Measurement System Units	Currency USD, EUR, VND, KRW, CNY, JPY, ...	Length Systems Kilometer (Metric), Mile (Imperial), Cape Foot (South African), Wa (Thai), Li (Chinese), ...	Weight Systems Kilogram (Metric), Pound (Imperial) Jin (Chinese), Geun (Korean), Catty/Kati (Malaysian), ...
Number of Systems	112	10	10
Total Samples (Facts × Systems)	16576	31200 (Kilometer), 7800 (Mile)	24130 (Kilogram), 3690 (Pound)
Number of Facts Example Question	148 *What is GDP per capita for Australia in 2021 in **JPY**?*	3120 (Kilometer), 780 (Mile) *What is the distance between New York and Los Angeles in **Li (Chinese)**?*	2413 (Kilogram), 369 (Pound) *What is the price of one **Korean Geun** of steak (beef) in Nigeria, in Nigerian Naira (NGN)?*

Table 1: **Measurement System Datasets:** Measurement system and units, as well as sample size by dataset type. For further information about each measurement system, see Appendix Table 5.

Code and dataset are available at `https://github.com/MinhDucBui/MeasurementSystemBias`.

Fiscal Data. To analyze currency systems, we collect international fiscal data focused on GDP per capita. We obtain the figures for 148 countries from the World Bank (The World Bank, 2024) and specifically use 2021 data to align with the models' knowledge cutoff, while at the same time minimizing recency biases (Moayeri et al., 2024). We obtain the average monthly exchange rates for 2021 from the International Monetary Fund to convert these fiscal values into other currencies (International Monetary Fund, 2024). To mitigate the impact of exchange rate fluctuations, we select the most favorable rate within that year, ensuring optimal performance. Note that fiscal data is typically reported in USD (Krugman, 1984).

Food Prices per Weight: Kilogram System. We differentiate between countries that use the "Kilogram" (metric) system and the "Pound" (imperial) system. In countries that use the "Kilogram" system, we collect food commodity prices for items such as maize, rice, beans, and sugar reported per kilogram (e.g., price of 1 kg of rice), expressed in the respective local currency. This data stems from World Food Program via HDX (World Food Program, 2025), covers 76 countries, and spans from 2010 to 2021.

Food Prices per Weight: Pound System. We obtain food price data for United States from the Bureau of Labor Statistics (U.S. Bureau of Labor Statistics, 2025). This dataset includes retail prices for a wide range of items, including fruits, meats, and other commodities, expressed in pound (e.g., the price of 1 pound of chicken).

City Distances: Kilometer System. We begin by selecting four countries in which the "Kilometer" (metric) system is the primary measurement system: Germany, Russia, China and Japan. For each country, we randomly select 40 cities from among its 80 largest cities. We then compute the pairwise distances between all selected cities within each country, resulting in 3,120 distance measurements in total. The shortest straight-line distances between cities were calculated using the Haversine formula, with city coordinates sourced from SimpleMaps (2025).

City Distances: Mile System. For the United States, where the "Mile" (imperial) system applies, we likewise compile a list of 80 largest cities by population and calculate the shortest road distances as for the "Kilometer" system.

3 Stability across Measurement Systems

We begin our investigation of RQ1 by analyzing what measurement systems LLMs default to when prompted without any instruction on the use of measurement. Next, we investigate RQ2 by examining whether LLM behavior shifts when instructed to use a measurement system different from its default and whether this affects performance.

3.1 Experimental Setup

Models. We test instruct-finetuned LLMs from different model families and of varying sizes. Specifically, we probe Qwen2.5 (72B, 7B) (Qwen et al., 2025), Llama 3.3 (72B), Llama 3.1 (72B, 8B) (Grattafiori et al., 2024) and Aya Expanse (32B, 8B) (Dang et al., 2024). We group the models by

System	Large	Small	Avg.
Currency (Fiscal Data)			
Local Currency	18%	16%	17%
Other			
= USD	**76%**	**82%**	**80%**
= Rest	<u>6%</u>	<u>2%</u>	<u>3%</u>
Weight Systems (Food Prices)			
Data: Kilogram			
Kilogram	**100%**	**100%**	**100%**
Pound	0%	0%	0%
Geun, Jin, ...	0%	0%	0%
Data: Pound			
Kilogram	0%	0%	0%
Pound	**100%**	**100%**	**100%**
Geun, Jin, ...	0%	0%	0%
Length Systems (City Distances)			
Data: Kilometer			
Kilometer	**100%**	**100%**	**100%**
Mile	0%	0%	0%
Cape Foot, Wa, ...	0%	0%	0%
Data: Mile			
Kilometer	37%	22%	27%
Mile	**63%**	**78%**	**73%**
Cape Foot, Wa, ...	0%	0%	0%

Table 2: **LLMs' Default Measurement Systems:** We report the total percentage of measurement systems that LLMs default to when given the freedom to choose.

their size into 'Large' (>=70B) and 'Small' (<70B). For more information, see Appendix B.2.

Prompts. To first isolate the effect of different measurement systems, we use English prompts in our experiments, ruling out the language of the prompt as a confounding factor. In §3.3, we also explore multilingual prompts. For robustness, we generate three prompt variants for each task. In each variant, the model is instructed to adhere to a specific measurement system, as illustrated by the following example:

> *Estimate the GDP per capita for Germany in the year 2021, **expressed in the currency ERN**. Your answer must only include the GDP per capita [...]*

We refer to Appendix B.1 for further details. Additionally, we specify the original name and country of origin for certain length and weight systems.

3.2 RQ1: Default Systems

We first analyze which measurement system the model defaults to. To that end, we let the model generate its response without the instruction to output the fact in a specific measurement.

Results. Table 2 summarizes the findings on default systems. For fiscal data, LLMs predominantly

default to using USD (80% on average), and almost exclusively rely on either USD or local currency (17% on average). For weight measurements, LLMs strictly employ either the imperial or metric system, in line with the data context. In the case of length data, distances in Germany, Japan, Russia and China are reported solely in the metric system, whereas, for city distances in the USA, 73% are expressed in the imperial system.

Discussion. Our observations suggest that **LLMs tend to adopt measurement systems that mirror the conventions present in the data, often defaulting to Western standards**, such as using USD for fiscal figures. This trend seems to reflect the broader influence of the West, emerging not through any direct global mandate but through market dynamics shaped by Western forces[2] (Krugman, 1984). Interestingly, Large and Small LLMs exhibit very similar behavior, with US city distances as the only (partial) exception: Small LLMs default more consistently (+15%) to the American cultural context, i.e., the imperial system.

3.3 RQ2: Disparity & Performance across Measurement Systems

We evaluate the stability of LLMs' responses across measurement systems by explicitly instructing them to produce outputs in a specified system. For each dataset, the default system is the one which LLMs used most frequently in §3.2 (Table 2). We then use the performance with the default system as a baseline for the alternative systems.

Evaluation Metric. We evaluate a model's performance by comparing its predictions to the ground truth. Because the magnitude of answers varies considerably, we resort to an inverse mean absolute percentage deviation (MAPD). To mitigate the impact of extreme predictions—which LLMs may occasionally produce—we cap the deviation at a maximum of 100%. The final metric is defined as follows:

$$APD_i^k = \min\left(\left|\frac{p(x_i, f^k)^z - y_i^z}{y_i^z}\right|, 1\right),$$

$$\text{MAPD}^k = 100 \times \left(1 - \frac{1}{N}\sum_{i=1}^{N} APD_i^k\right).$$

---

[2]For example, before World War I, the pound sterling was the leading international currency; later, the dollar and the pound shared this role, and eventually the dollar dominated during the Bretton Woods era.
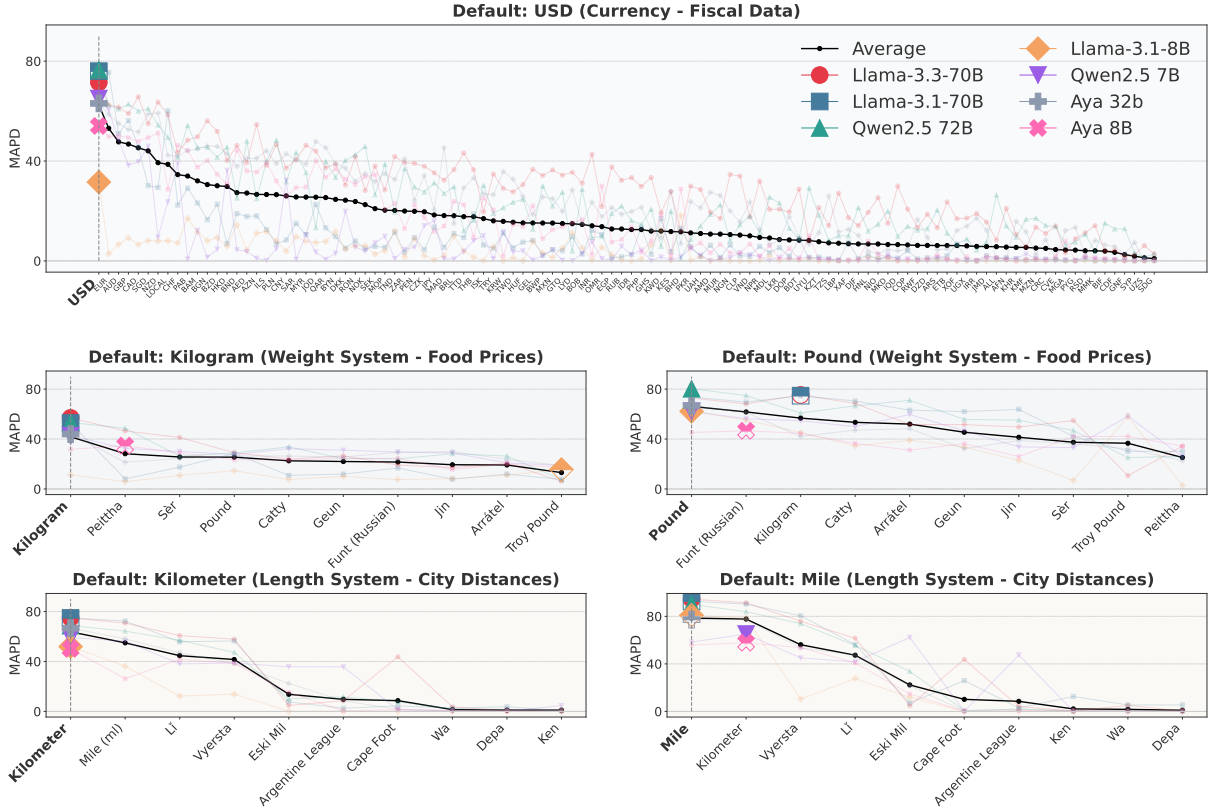
Figure 2: **Disparity and Performance of LLMs Across Different Measurement Systems:** We report the performance in terms of MAPD (cf. §3.3). The best-performing system (higher is better) for each model is highlighted with bold points. Fully filled points indicate statistically significant improvement over the second-best system, whereas half-filled points do not. We report the raw numbers and their significance in Appendix Table 6.

Here, $p(x_i, f^k)$ is the model's prediction for sample i using system k, and $p(x_i, f^k)^z$ represents its conversion to measurement system z. The ground truth value for sample i, also expressed in system z, is denoted by y_i^z. We note that higher values indicate better performance. An inverse MAPD of 76.0 indicates *approximately* that the model's estimates deviate from the ground truth by 24%.

We test whether the behavior of two systems differs significantly by comparing their paired absolute percentage deviation distributions over samples. We employ the non-parametric Wilcoxon Signed-Rank Test (Wilcoxon, 1945), which establishes whether the median difference between the paired values significantly deviates from zero. We use a high significance threshold of $p = 0.001$.

Results. Table 2 summarizes our findings. On average across all models, the default system consistently delivers the best performance. Among 35 model–dataset combinations, it outperforms all others in 28 cases, with statistically significant gains over the second-best system in 22 of those. For example, in the fiscal dataset, using USD yields

the highest performance for *every* single model (average MAPD of 63%). Similarly, in the weight and length system dataset, models overwhelmingly yield optimal results with the default system. To illustrate, Llama 3.3 70B achieves a 57% MAPD with the default "Kilogram" system, but its performance drops significantly to 29% when using the "Pound" system. In the length dataset, Llama 3.3 70B achieves a 75% MAPD on cities with the default "Kilometer" system and shows a 4-point performance drop when evaluated with the "Mile" system. Only for US cities, we observe a smaller average difference from the "Kilometer" system and with Qwen2.5 7B being the only model to show a statistically significant improvement when using this alternative system. While other models also perform better with certain alternative systems, their improvements are not statistically significant. Finally, we observe uneven performance drops across alternative measurement systems, as reflected in the drop of average scores.

Discussion. In sum, we find that **current LLMs are not stable across different measurement sys-**
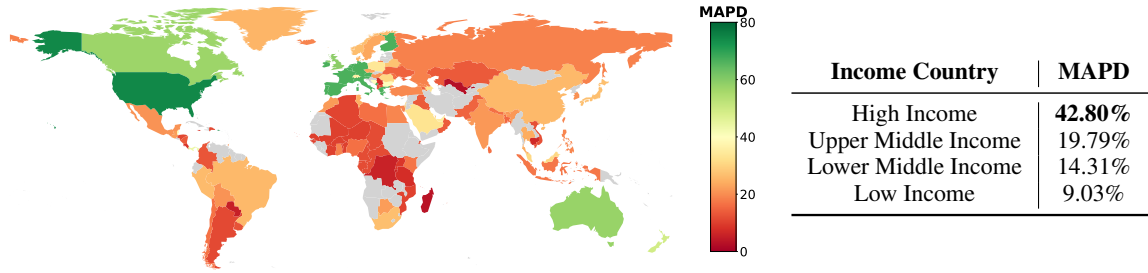
Figure 3: **Currency Performance Mapped and Categorized by Income Groups:** We visualize the performance for all currencies—averaged across large models (over 70B parameters)—on a world map, with countries categorized by income levels.

tems; LLMs yield the best performance when used with their default system and generally suffer significant and pronounced drops in performance for alternative systems. Such issues have severe implications for users from marginalized cultural backgrounds, who stand to benefit the most from equitable access to language technology. When LLMs misrepresent or inaccurately convert measurements, e.g., wrongly state the value of consumer goods in local currency, they alienate users by failing to correctly adapt to their cultural context.

Analysis: Currency Bias. We visualize currency performance of our biggest LLMs (>=70B) on a world map, associating each country with its respective currency. Additionally, we group countries by income levels provided by the World Bank (Fantom and Serajuddin, 2016). We immediately observe obvious disparities across income groups. LLMs achieve an average performance of 43% with currencies from high-income countries, which is an impressive 34% better than with currencies from low-income countries. This constitutes a clear bias that favors the richest (i.e., their currencies) and affects the poorest. As already highlighted in §3.3, using non-default systems adversely affects LLMs' performance: this analysis **reveals that this effect disproportionately disadvantages low-income countries**.

Analysis: Multilingual Setting. We further examine whether language affects our findings by aligning the prompt language with the measurement system used. For instance, when using the Korean currency, we analyze the performance gap between the Korean currency and USD (default measurement system) under Korean prompt language. For each dataset, we select two languages from among Turkish (TR), Korean (KR), Chinese (ZH), and Japanese

Setting	Llama 3.3 70B	Qwen2.5 72B	Aya 32B
Currency (Fiscal Data)			
Default=USD			
Δ_{KR} (KRW)	-37.89*	-27.43*	-36.36*
Δ_{TR} (TRY)	-57.01*	-42.22*	-49.13*
Weight System (Food Prices)			
Default=Kilogram			
Δ_{KR} (Geun)	-21.99*	-31.12*	-22.11*
Δ_{ZH} (Jin)	-25.39*	-26.79*	-11.33*
Default=Pound			
Δ_{KR} (Geun)	-3.70	-12.54*	-39.14*
Δ_{ZH} (Jin)	-1.31	-20.25*	-2.29
Length System (City Distances)			
Default=Kilometer			
Δ_{JA} (Ken)	-80.09*	-75.71*	-73.90*
Δ_{ZH} (Lǐ)	-20.19*	-13.18*	-15.03*
Default=Mile			
Δ_{JA} (Ken)	-89.68*	-86.97*	-84.98*
Δ_{ZH} (Lǐ)	-30.33*	-35.26*	-35.45*

Table 3: **Aligning Language with Measurement System:** We compare performance between the default system and the system aligned with the prompt language (e.g., KRW vs. USD for Korean prompts). The language is shown as a subscript, and the aligned system in brackets. Red cells mark significant differences.

(JA)—representing high- to medium-resource languages. For each language, a native speaker reviews our prompts to ensure correctness. We report the results in Table 3.

In nearly all cases, we observe a significant performance drop. For example, even when using the Korean currency with Korean prompt language, Llama 3.3 70B still experiences a significant decrease of 38 points. The only cases without a significant drop are Llama 3.3 70B in the weight task using the "Pound" system and Aya 32B when evaluating the Chinese "Jin" unit, also under the "Pound" system. We conclude that even when the language

is aligned, significant drops persist in nearly all cases, highlighting the **significant instability of measurement systems independent of language in almost all cases**.

Figure 4: **Visualization of Multi-Hop Reasoning:** Answering w.r.t. an alternative measurement system requires multi-hop reasoning, consisting of the following steps: (i) retrieving the answer in the default system, (ii) retrieving the conversion rate between the default and alternative system, and (iii) combine the results of (i) and (ii) with arithmetic reasoning.

4 RQ3: Multi-Hop Reasoning

We first explain how answering questions w.r.t. alternative measurement systems represents a multi-hop reasoning problem and then explore whether the LLMs' performance changes with (i) a sequence of prompts that aim at individual reasoning steps and (ii) with chain-of-thought prompts, i.e., by explicitly asking the LLMs to provide the explanation for their final answer.

Multi-Hop Questions. We adopt the definition and notation from Mavi et al. (2024): A multi-hop question (MHQ) is a question that cannot be answered using a single document or passage (context); instead, it requires combining information from multiple sources. For example, a question like "What is the capital of the country whose currency has the highest exchange rate against the US dollar?" necessitates gathering and synthesizing information from several contexts.

Formally, let $\mathbb{C}$ denote the set of all contexts, $\mathcal{Q}$ the set of all questions, and $\mathcal{A}$ the set of all possible answers. For a question $q \in \mathcal{Q}$ and a subset of related contexts $C \subseteq \mathbb{C}$, the task is to approximate a function

$$f : \mathcal{Q} \times 2^{\mathbb{C}} \to \mathcal{A} \cup \{\Phi\},$$

such that

$$f(q, C) = \begin{cases} a \in \mathcal{A} & \text{if there exists a subset} \\ & P_q = \{p_1, \dots, p_k\} \subseteq C, \\ & \text{with } k > 1, \text{such} \\ & \text{that } P_q \models (a \text{ answers } q), \\ \Phi & \text{otherwise.} \end{cases}$$

where $\models$ denotes entailment and Φ is returned when q is unanswerable using C. The requirement that $k > 1$ ensures that the question is not solvable in a single hop (i.e., with $k = 1$, the task reduces to traditional QA).

When we ask questions of the form "*What is <Fact> in <System>?*", where *<Fact>* represents information inherently associated with a particular measurement system (e.g., GDP is intrinsically tied to currency) and the system is not the default one for the model, we are effectively asking the model to solve a multi-hop reasoning task consisting of the following steps:

1. Source Fact Retrieval: Extract the fact from its original measurement system (the source system, S_{src}). This constitutes a single-hop operation using one context, p_1.

2. Conversion Retrieval: Obtain the information required to convert the fact from the source system S_{src} to the target system S_{trg}. This is another single-hop operation that relies on a different context, p_2.

3. Arithmetic Reasoning: Apply the conversion information to the retrieved fact. This requires arithmetic reasoning and may require one or more additional contexts $\{p_3, \dots\}$.

In this scenario, the model must integrate multiple contexts $\{p1, p2, p3, ...\}$ to arrive at an answer, which aligns with our multi-hop definition. Figure 4 provides a visual overview of this process.

Reasoning Chains. Recognizing measurement system questions as multi-hop reasoning problems, we next investigate two strategies that push LLMs to reason: (1) *Sequential Single Hops* explicitly breaks down the problem for the LLM into individual reasoning steps and (2) *Chain-of-Thought (CoT)* prompts implicitly nudge the LLMs to reason by requiring an explanation for the answer.

Sequential Single Hops (Seq). We first ask the LLMs to solve individual hops sequentially. To this end, we structure single-hop queries as illustrated in Figure 4. We feed the hops to the LLM in sequence, integrating the answer from the previous step into the next prompt: e.g., after obtaining the answer to "*What is the exchange rate from USD to*

Small Models				Large Models					
Avg. Performance				Avg. Performance				Avg. Cost	
Default Sys.	Alternative Systems			Default Sys.	Alternative Systems			Alternative Sys.	
NoReas.	ΔNoReas.	ΔSeq.	ΔCoT	NoReas.	ΔNoReas.	ΔSeq.	ΔCoT	ΔSeq.	ΔCoT
Default: USD (Currency Systems)									
53.43	-77.45%	-31.32%	-44.15%	74.67	-72.45%	-0.12%	+0.79%	+189%	+302%
Default: Kilogram (Weight Systems)									
32.43	-33.11%	-17.92%	-13.88%	54.33	-59.13%	-4.82%	-8.05%	+191%	+206%
Default: Pound (Weight Systems)									
59.08	-33.97%	-19.10%	-33.31%	75.61	-29.25%	-4.72%	-9.39%	+199%	+187%
Default: Kilometer (Length Systems)									
57.00	-63.64%	-31.82%	-41.74%	72.84	-57.93%	-7.88%	-3.98%	+180%	+210%
Default: Mile (Length Systems)									
68.13	-74.36%	-34.55%	-49.95%	92.46	-73.13%	-7.18%	-7.50%	+197%	+185%

Table 4: **Performance and Cost Increases across Reasoning Strategies:** *Performance:* We report the percentage change relative to the default system (*Default Sys.*) without reasoning (*NoReas.*, cf. §3.3). We highlight cells in red when the performance drop exceeds 10%. *Cost Increase:* We report the additional cost incurred compared to using no reasoning.

VND in 2021?" we incorporate it into the subsequent query: *"Convert <number> USD to VND."*

Chain-of-Thought (CoT). We explicitly prompt the LLM to detail its intermediate steps leading to the final answer. Specifically, we prompt the model with: *"What is the GDP of Germany in ERN? Explain first. Be short."*

Evaluation. We report the change in MAPD relative to the default measurement system prompted without any reasoning (as discussed in Section 3.3), as shown in Table 4. Additionally, we examine the cost associated with these reasoning methods. For this purpose, we refer to the pricing provided by an LLM API provider[3] for the LLaMA 3.3 70B family and report the cost increase relative to the no-reasoning baseline (right part of Table 4). We further examine which measurement systems, in addition to the targeted system, are referenced in the CoT reasoning (see Figure 5).

Results. Small LLMs experience a significant performance drop and show considerable instability, often failing to reason effectively, whereas Large models maintain stability across different measurement systems, with only slight performance deviations from the baseline. However, the cost of reasoning with Large models is considerably higher compared to the no-reasoning baseline. Sequential incur a cost up to 199% of the cost without reasoning, whereas CoT on average increases the test-time compute cost to 300% for currency conversions and around 200% for weight and length system conversions.

Discussion. Prompting the model, explicitly (Seq) or implicitly (CoT) to reason does not benefit Small models. For Large models, reasoning does stabilize the performance to the level of the default measurement system without reasoning. This improvement comes at a significant cost: using reasoning increases test-time compute and, in turn, expenses by 180% to 300%, depending on the dimension of cultural context. While reasoning is an effective solution, its **increased cost disproportionately affects users who rely on alternative measurement systems rather than the LLM's default**. This discrepancy arises as models triangulate the conversion process step-by-step from the default measurement system. When faced with non-default measurement systems, **LLMs engage in additional reasoning steps to convert the result from the default system** (e.g., LLMs start from USD for currency conversion in 91% of the cases; see Figure 5), which demands additional computation. Consequently, users with cultural contexts that are misaligned with the LLMs' default (Anglo-centric context) face higher costs—this is yet another obstacle that disproportionately hinders access to state-of-the-art language technology for the least developed parts of the world.
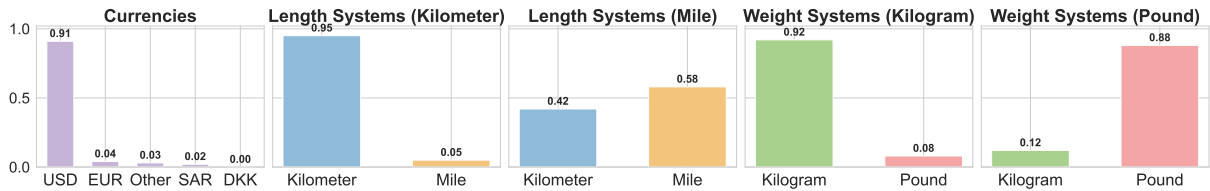
[3]https://groq.com/pricing/

Figure 5: **Frequency of Measurement Systems as Starting Points in CoT Reasoning.** We count how often each measurement system is chosen as the starting point for the model's CoT reasoning and report the proportion. We average this proportion across all models.

5 Related Work

Biases in Factual Retrieval. Recent research has examined LLMs' factual recall (Mallen et al., 2023; Kandpal et al., 2023; Sun et al., 2024; Tang et al., 2023, *inter alia*). While these models often provide correct answers, they can hallucinate (Huang et al., 2025). Studies also show that recall accuracy correlates with an entity's popularity—with higher error rates for lesser-known subjects (Mallen et al., 2023; Sun et al., 2024)—and reveal geographical biases along regional and income lines (Moayeri et al., 2024; Manvi et al., 2024).

In contrast to prior work, our study shifts attention to the systems representing facts, noting that while measurement systems may influence depiction, the underlying facts remain unchanged. Park et al. (2022) demonstrate that small encoder models struggle with comparisons across different measurement systems. Additionally, Dinu et al. (2020) explicitly incorporates measurement systems into translation models. However, unlike these studies, our work focuses on decoder-based LLMs, emphasizing factual retrieval rather than simple comparisons or translations.

Cultural Biases in LLMs. Kirk et al. (2024) underscore the disparities in how LLMs are used across diverse cultural contexts. However, numerous studies have quantified and identified cultural biases in LLMs and their multimodal extensions (Cao et al., 2024; Shi et al., 2024; Myung et al., 2024; Bui et al., 2024, *inter alia*). The consensus is that LLMs tend to align more closely with Western values, which limits their effectiveness in cross-cultural contexts. For instance, Lee et al. (2024) demonstrate that current LLMs achieve higher accuracies on hate speech labels derived from the Anglosphere. In our work, we further show that the same pattern models also favor measurement systems predominantly used in the Western world.

Multi-Hop Reasoning. Research on multi-hop rea-

soning has produced numerous datasets and analyses (Yang et al., 2018; Fang et al., 2020; Chen et al., 2020; Trivedi et al., 2022, *inter alia*). Recent work has shifted toward implicit multi-hop questions that require models to recall and integrate facts without explicit reasoning steps (Press et al., 2023; Biran et al., 2024; Li et al., 2024; Yang et al., 2024). Although LLMs can connect internal knowledge, they still struggle when key reasoning details are absent (Press et al., 2023; Dziri et al., 2023). Moreover, while prior work mainly examines concatenated, independent facts, our analysis shows that natural multi-hop questions from measurement systems present unique challenges.

6 Conclusion

In this work, we show that LLMs exhibit significant variability in performance based on the measurement system embedded in the prompt for factual retrieval. Specifically, performance declines when the model encounters non-default measurement systems. We show that reasoning (e.g., CoT) improves the model's ability to handle alternative systems questions but it leads to an increase in test-time compute, which disproportionately affects users whose cultural contexts are misaligned with the LLMs' defaults.

Limitations

In this study, we first prompt LLMs in English to examine how measurement systems affect factual retrieval accuracy. However, the interplay between language and measurement systems and its impact on performance remains an important avenue for future research. Users from underrepresented communities are much more likely to interact with dialogue systems in their native languages. Nonetheless, investigating how well LLMs generalize across measurement systems in English represents a 'best-case' scenario, given that these models perform best in English. Consequently, our core

results would be even more pronounced in other languages, where LLMs are likely to mistranslate measurements more frequently and significantly, and incur higher inference costs due to tokenization into more subwords.

Similarly, while our investigation examines a selected subset of measurement systems and units, we acknowledge the challenges of accessing comprehensive data—especially regarding systems still actively used within marginalized groups. Despite these constraints, our work successfully highlights key dynamics and interactions. We view this as a promising starting point that opens up opportunities for expanding the scope of research to include a broader array of measurement systems.

Finally, we acknowledge that there are more effective tools for retrieving facts and performing unit conversions than solely relying on LLMs, such as RAG systems or dedicated conversion tools. However, we emphasize that users are more inclined to interact with a simple LLM chatbot for these seemingly straightforward tasks rather than resorting to more complex tools.

Ethical Statement

The sources of our data, including the World Bank, IMF, and HDX, may reflect biases inherent in data collection methodologies, economic models, or reporting structures. These biases could influence the outcomes of analyses based on the dataset. Any model trained or analyzed solely on our dataset may inherit assumptions and limitations from the data sources.

We use AI assistants, specifically o3-mini, to help edit sentences in our paper writing.

Acknowledgments

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. *Preprint*, arXiv:2309.14316.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130, Miami, Florida, USA. Association for Computational Linguistics.

Minh Duc Bui, Katharina von der Wense, and Anne Lauscher. 2024. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. *Preprint*, arXiv:2411.03888.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Georgiana Dinu, Prashant Mathur, Marcello Federico, Stanislas Lauly, and Yaser Al-Onaizan. 2020. Joint translation and unit conversion for end-to-end localization. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 265–271, Online. Association for Computational Linguistics.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: limits of transformers on compositionality. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Neil Fantom and Umar Serajuddin. 2016. The world bank's classification of countries by income. Policy Research Working Paper 7528, World Bank. License: CC BY 3.0 IGO.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in*

Natural Language Processing, pages 12216–12235, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

International Monetary Fund. 2024. International Financial Statistics: Exchange Rates. Accessed: 2024-02-13.

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.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Paul R. Krugman. 1984. The international role of the dollar: Theory and prospect. In *Exchange Rate Theory and Practice*, pages 261–278. National Bureau of Economic Research (NBER).

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. Understanding and patching compositional reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9668–9688, Bangkok, Thailand. 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.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering. *Preprint*, arXiv:2204.09140.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1211–1228, New York, NY, USA. Association for Computing Machinery.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen H Muhammad, Kiwoong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Sungjin Park, Seungwoo Ryu, and Edward Choi. 2022. Do language models understand measurements? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1782–1792, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

SimpleMaps. 2025. World cities database. https://simplemaps.com/data/world-cities. Accessed: 2025-05-28.

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.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

The World Bank. 2024. GDP per capita (current US). Accessed: 2024-02-13.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

U.S. Bureau of Labor Statistics. 2025. Consumer price index (cpi). Accessed: 2025-02-13.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

World Food Program. 2025. Consumer price index (cpi). Accessed: 2025-02-13.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Dataset

A.1 Measurement Systems

The exhaustive list of measures used in the dataset are shown in Table 7.

For ease of understanding, we additionally summarize the usage context and prevalence of each system in Table 5.

A.2 Terms of Use

Our research is conducted in the public interest under GDPR, meeting the criteria for substantial academic research. To respect all data providers, our published work is licensed under CC BY-NC-ND 4.0. We use datasets with the following licenses:

World Bank Group: The fiscal data is provided under an Access to Information policy and distributed as CC-BY 4.0.

Exchange Rates: Data from the IMF is available for download, extraction, and use under its open terms.

Food Prices: HDX and World Food Programme data use the Creative Commons Attribution for Intergovernmental Organisations license, while U.S. food prices from the Bureau of Labor Statistics are in the public domain.

City Distances: Data from the Simplemaps domain is included in approximate form for reproducibility, with proper attribution as required by their terms.

B Experimental Setup

B.1 Prompts

We report all prompts used in this paper.

Base Prompts (No Reasoning). We create three base prompt variants for the no reasoning part:

- Estimate the <MEASUREMENT> in year <YEAR>. Your answer must only include the <MEASUREMENT> as a decimal number without any abbreviations and explanation. Final number:

- Approximate the <MEASUREMENT> in year <YEAR>. Your answer must only include the <MEASUREMENT> as a decimal

System	Usage Description	Usage Level
Currencies		
ALL	Currencies are used in every day life	*High*
Weight System		
Kilogram	Official SI base unit of mass	*High*
Pound	Imperial/US customary unit	*High*
Geun (Korea)	Korean traditional unit; still used in markets for meat.	*Medium*
Arrátel (Portuguese)	Historical Portuguese unit; obsolete.	*Low*
Funt (Russia)	Used historically in local trade and marketplaces; obsolete.	*Low*
Peittha	Burmese traditional unit, still used in Myanmar for local trade.	*Medium*
Troy pound	Specialized unit used for precious metals like gold and silver.	*Medium*
Jin (Mainland China)	Used in markets, and for measuring gold, silver and Chinese medicines.	*Medium*
Catty/Kati (Malaysian)	Traditional weight still used in some local markets in Southeast Asia.	*Low*
Sèr (Pakistani)	Traditional unit used historically in the Indian subcontinent; mostly obsolete.	*Low*
Length Systems		
Kilometer	Used in most countries as the official system.	*High*
Mile	Used primarily in the US for everyday measurements.	*High*
Li (China)	Traditional Chinese unit of length, historically used in land measurement.	*Low*
Ken (Japan)	Traditional Japanese unit of length, commonly used in architecture.	*Medium*
Eski Mil (Turkish)	Ottoman-era unit, obsolete in modern Turkey.	*Low*
Vyersta (Russian)	Historical Russian unit; obsolete.	*Low*
Cape Foot (South Africa)	Colonial-era unit, obsolete in contemporary use.	*Low*
Depa (Indonesian)	Traditional unit based on armspan, now rarely used.	*Low*
Argentine league (legua)	Historical unit used for long distances in rural Argentina; obsolete.	*Low*
Wa (Thai)	Traditional Thai unit still used for land measurement.	*Medium*

Table 5: Overview of diverse measurement systems categorized by type (Currency, Weight, and Length), including their usage descriptions and estimated level of usage. We indicate *High* for systems that are officially used in the country, *Medium* for those that are not official but still used in specific local contexts, and *Low* for systems that are largely obsolete in everyday usage and are considered historical units.

```
number without any abbreviations and
explanation. Final number:
```

- ```
  Guess the <MEASUREMENT> in the year
  <YEAR>. Your answer must only include
  the <MEASUREMENT> as a decimal
  number without any abbreviations and
  explanation. Final number:
  ```

Base Prompts (CoT). To enable Chain-of-Thought (CoT) reasoning, we modify each base prompt. Specifically, we replace the instruction `"Your answer must only include the <MEASUREMENT> as a decimal number without any abbreviations and explanation. Final number:"` with `"Explain first and then give your final answer with 'Answer: <Your answer in decimal number>'. Be short:"`

Variable Definitions. For different measurement systems, the variable `<MEASUREMENT>` is defined as follows:

- **Currency:** `<MEASUREMENT>` = GDP per capita of `<COUNTRY>` in the currency `<CURRENCY>`

- **Weight Systems:** `<MEASUREMENT>` = the price of `<FOOD>` per `<WEIGHTSYSTEM>` in `<COUNTRY>`

- **Length System:** `<MEASUREMENT>` = distance between cities `<CITY1>` and `<CITY2>` in `<COUNTRY>` expressed in `<LENGTH SYSTEM>`

B.2 Model Details

We employ the Hugging Face implementation for all models, which are run on two A100 80GB GPUs. Outputs are generated with a limit of 40 new tokens per request, and each large LLM processes a dataset in approximately 1.5 hours with a batch size of 64. For CoT prompting, we increase the token limit, which extends the processing time to nearly 24 hours.

	Llama 3.3	Llama 3.1		Qwen2.5		Aya		Average
	70B	**70B**	**8B**	**72B**	**7B**	**32B**	**8B**	
Currency (Fiscal Data)								
Default = USD	*71.51**	*75.99*	*31.58**	*76.51**	*65.01**	*63.08**	*54.05**	*62.53**
EUR	62.45	75.48	2.84	62.34	60.34	58.55	49.79	53.11
AUD	61.55	50.99	6.66	60.88	48.34	55.08	50.08	47.65
GBP	59.00	56.24	9.20	62.76	38.32	52.67	49.40	46.80
CAD	65.73	45.59	6.57	60.00	39.68	51.63	47.87	45.30
SGD	54.99	30.16	8.14	60.94	45.94	58.84	49.40	44.06
NZD	63.59	29.40	7.96	54.41	9.48	59.24	51.18	39.32
LOCAL	58.55	60.13	8.08	49.49	23.67	42.27	28.64	38.69
...	(all)	(all)	(all)	(all)	(all)	(all)	(all)	...
UZS	6.15	0.38	0.00	1.44	0.61	0.13	0.01	1.25
SDG	2.76	0.18	0.11	2.88	0.10	0.17	0.17	0.91
Weight System (Food Prices)								
Default=Kilogram	*56.94**	*52.83**	*11.33*	*53.22**	*42.67**	*43.99**	*31.74*	*41.82*
Peittha	46.48	8.05	5.82	48.69	32.43	21.46	**34.73**	28.24
Sèr	41.28	17.49	10.85	26.31	30.10	24.44	28.81	25.61
Pound	29.04	28.44	14.68	28.81	27.56	26.43	23.40	25.48
Catty	23.95	10.90	7.57	33.71	32.27	26.40	23.13	22.56
Geun	26.52	11.91	10.04	24.36	31.05	25.54	24.58	22.00
Funt (Russian)	19.19	16.79	7.51	24.06	29.57	29.39	24.48	21.57
Jin	17.16	7.86	8.29	28.27	28.85	29.88	15.57	19.41
Arrátel	20.55	11.88	11.04	26.39	20.83	23.74	20.03	19.21
Troy Pound	6.21	7.61	**15.56**	7.58	18.97	18.77	17.09	13.11
Default=Pound	*72.85*	*73.68*	*62.20**	*80.29**	*62.27**	*66.61**	*45.25*	*66.16*
Funt (Russian)	68.15	69.61	55.22	74.82	56.25	61.25	**46.61**	61.70
Kilogram	**75.32**	**74.74**	45.12	60.82	54.66	42.11	44.56	56.76
Catty	68.54	70.46	34.57	66.56	49.91	47.11	36.49	53.38
Arrátel	51.63	63.14	38.99	71.05	59.68	48.24	31.09	51.98
Geun	51.59	62.01	33.56	55.71	46.59	32.37	36.24	45.44
Jin	49.62	63.81	22.85	55.12	33.75	38.81	25.99	41.42
Sèr	54.83	42.14	6.95	46.84	33.07	36.11	42.07	37.43
Troy Pound	10.70	30.20	57.83	25.12	58.28	32.14	42.14	36.63
Peittha	34.35	30.16	2.77	26.23	25.88	22.74	34.31	25.21
Length System (City Distances)								
Default=Kilometer	*74.92**	*74.85*	*51.95**	*68.74**	*59.26*	*66.66**	*50.14**	*63.79*
Mile	70.81	72.42	36.25	64.39	56.25	58.14	26.27	54.93
Lǐ	60.75	55.83	12.30	56.92	38.05	46.94	41.76	44.65
Vyersta	57.88	56.47	13.88	47.28	38.80	38.33	38.55	41.60
Eski Mil	4.78	7.64	0.11	10.07	35.81	22.41	15.24	13.72
Argentine League	8.58	2.55	0.20	11.71	35.71	8.60	0.41	9.68
Cape Foot	43.77	4.41	0.00	7.79	1.37	1.61	1.51	8.64
Wa	3.65	2.72	0.10	2.03	0.42	0.37	0.79	1.44
Depa	0.32	3.77	0.14	1.71	0.23	0.50	0.99	1.10
Ken	0.27	0.70	0.07	0.60	4.72	0.27	0.67	1.04
Default=Mile	*94.54**	*92.72*	*80.96*	*90.11**	*58.40*	*77.25*	*55.92*	*78.56*
Kilometer	91.35	90.31	78.40	83.88	**65.10**	77.24	**57.84**	77.73
Vyersta	75.83	80.50	10.43	73.97	45.05	53.45	53.66	56.13
Lǐ	61.57	56.29	27.66	55.53	41.82	47.46	40.84	47.31
Eski Mil	4.78	7.12	11.84	33.70	62.27	21.59	14.74	22.29
Cape Foot	43.72	25.95	0.01	0.84	0.12	0.39	0.10	10.16
Argentine League	4.24	2.89	0.34	1.87	47.48	2.02	0.61	8.49
Ken	0.27	12.56	0.03	0.26	0.15	0.76	0.75	2.11
Wa	5.15	5.48	0.19	0.52	0.17	0.31	0.27	1.73
Depa	0.22	5.46	0.14	0.26	0.12	0.40	0.38	1.00

Table 6: **Disparity and Performance of LLMs Across Different Measurement Systems:** The first row (shaded in gray) within each group represents the model's default measurement system. We report the performance in terms of MAPD (cf. §3.3). The best performance in each column (higher is better) are shown in **bold**. *Disparity:* The best values are additionally marked with *, if statistically significant differences ($p < 0.001$) compared to the second best system exists. Cells highlighted in red indicate cases where a significant difference was observed.

Currency (Fiscal Data)					
Currency code	Description	Conversion (to USD)	Currency code	Description	Conversion (to USD)
AED	United Arab Emirates Dirham	3.6725	KMF	Comorian Franc	418.0677243
AFN	Afghan Afghani	77.41418667	KRW	South Korean Won	1149.108333
ALL	Albanian Lek	104.0547222	KWD	Kuwaiti Dinar	0.301641667
AMD	Armenian Dram	503.6105556	KZT	Kazakhstani Tenge	425.9513889
ARS	Argentine Peso	95.18694444	LBP	Lebanese Pound	1507.5
AUD	Australian Dollar	1.337372732	LKR	Sri Lankan Rupee	201.3890667
AZN	Azerbaijani Manat	1.7	LYD	Libyan Dinar	4.523094444
BAM	Bosnia-Herzegovina Convertible Mark	1.6673125	MDL	Moldovan Leu	17.66990278
			MAD	Moroccan Dirham	9.045236111
BDT	Bangladeshi Taka	85.24412794	MGA	Malagasy Ariary	3854.072778
BGN	Bulgarian Lev	1.661877778	MKD	Macedonian Denar	52.55698889
BHD	Bahraini Dinar	0.376	MMK	Myanma Kyat	1610.540741
BIF	Burundian Franc	1978.428889	MOP	Macanese Pataca	8.008919444
BND	Brunei Dollar	1.345794444	MUR	Mauritian Rupee	41.72418333
BOB	Bolivian Boliviano	6.91	MXN	Mexican Peso	20.654375
BRL	Brazilian Real	5.368375	MYR	Malaysian Ringgit	4.143145833
BWP	Botswanan Pula	11.19945833	MZN	Mozambican Metical	66.0225
BYN	Belarusian Ruble	2.533175	NGN	Nigerian Naira	406.2341667
BZD	Belize Dollar	2	NIO	Nicaraguan Cordoba	35.20105833
CAD	Canadian Dollar	1.251083333	NOK	Norwegian Krone	8.631944444
CDF	Congolese Franc	1989.948356	NPR	Nepalese Rupee	118.7686111
CHF	Swiss Franc	0.9150375	NZD	New Zealand Dollar	1.414869444
CLP	Chilean Peso	779.4822222	OMR	Omani Rial	0.3845
CNY	Chinese Yuan	6.457772222	PAB	Panamanian Balboa	1
COP	Colombian Peso	3784.358056	PEN	Peruvian Nuevo Sol	3.892861111
CRC	Costa Rican Colon	625.7229167	PHP	Philippine Peso	49.43672222
CVE	Cape Verdean Escudo	93.765085	PKR	Pakistani Rupee	164.1562869
CZK	Czech Republic Koruna	21.67958333	PLN	Polish Zloty	3.889841667
DJF	Djiboutian Franc	177.721	PYG	Paraguayan Guarani	6676.519722
DKK	Danish Krone	6.317175	QAR	Qatari Rial	3.64
DOP	Dominican Peso	57.25499444	RON	Romanian Leu	4.192530556
DZD	Algerian Dinar	135.6026	RSD	Serbian Dinar	99.94511944
EGP	Egyptian Pound	15.65245278	RUB	Russian Ruble	73.53540833
ETB	Ethiopian Birr	44.25473611	RWF	Rwandan Franc	991.6896025
EUR	Euro	0.84978685	SAR	Saudi Riyal	3.75
GBP	British Pound Sterling	0.72827027	SDG	Sudanese Pound	299.4778056
GEL	Georgian Lari	3.25225	SEK	Swedish Krona	8.649225
GHS	Ghanaian Cedi	5.851411111	SGD	Singapore Dollar	1.345016667
GNF	Guinean Franc	9933.2195	SYP	Syrian Pound	1256
GTQ	Guatemalan Quetzal	7.7412875	THB	Thai Baht	32.01021944
HKD	Hong Kong Dollar	7.776777778	TND	Tunisian Dinar	2.805069444
HNL	Honduran Lempira	24.0325	TRY	Turkish Lira	9.808511111
HUF	Hungarian Forint	304.9938889	TTD	Trinidad and Tobago Dollar	6.747011111
IDR	Indonesian Rupiah	14333.63958	TWD	New Taiwan Dollar	28.05294444
ILS	Israeli New Sheqel	3.224777778	TZS	Tanzanian Shilling	2296.881064
INR	Indian Rupee	73.82430833	UAH	Ukrainian Hryvnia	27.27136389
IQD	Iraqi Dinar	1450	UGX	Ugandan Shilling	3601.599952
IRR	Iranian Rial	42000	USD	US Dollar	1
ISK	Icelandic Krona	125.9425	UYU	Uruguayan Peso	43.51508333
JMD	Jamaican Dollar	150.3607333	UZS	Uzbekistan Som	10646.02833
JOD	Jordanian Dinar	0.71	VND	Vietnamese Dong	23177.22222
JPY	Japanese Yen	109.6466667	XAF	CFA Franc BEAC	557.4236325
KES	Kenyan Shilling	110.2038889	XOF	CFA Franc BCEAO	557.4236325
KHR	Cambodian Riel	4097.125	ZAR	South African Rand	14.955069

Weight System					
Measure	Description	Conversion (in KG)	Measure	Description	Conversion (in KG)
Kilogram (kg)		1	Geun (Korea)		0.6
Pound (lb)		0.453592	Troy pound		0.37324
Arrátel (Portugese)		0.459	Funt (Russia)		0.409517
catty/kati (Malaysian)		0.60479	Sèr (80 Tolä, Pakistani)		0.93310
Jin (Mainland China)		0.5	Peittha (Burmamese)		1.63293

Length System					
Measure	Description	Conversion (in KM)	Measure	Description	Conversion (in KM)
Kilometer (km)		1	Mile (mi)		1.60934
Lǐ (China)		0.5	Eski Mil (Turkish)		1.89435
Vyersta (Russian)		1.0668	Ken (Japan)		0.001818
Cape Foot (South Africa)		0.0003148584	Depa (Indonesian)		0.0016
Argentine league (legua)		5.572	Wa (Thai)		0.002

Table 7: **Measurement System Description.** Description of measures used in the dataset.