THE LEADERBOARD ILLUSION: A RISK ANALYSIS OF PERFORMANCE DEGRADATION IN MULTILINGUAL LARGE LANGUAGE MODELS

Hejroe
Independent Researcher
United Kingdom
hejroe.ai@gmail.com

Gemini Pro 2.5
Google
United States of America

November 16, 2025

ABSTRACT

The selection of Large Language Models (LLMs) for deployment in enterprise and public sector applications is often guided by performance on English-centric benchmarks. This paper challenges the validity of this approach through a systematic, cross-lingual risk analysis of model reliability. We evaluate eleven prominent open-source LLMs, all runnable on consumer-grade hardware, on a corpus derived from the established MMLU and GSM8K benchmarks. Our fully automated methodology quantifies "performance drift" across UK English (EN), German (DE), and Spanish (ES), employing a novel Hybrid Automated Scoring protocol that uses semantic similarity to objectively assess procedural reasoning against a ground truth.

Our findings reveal a significant "leaderboard illusion": high performance in English is a dangerously poor predictor of a model's capabilities in other languages. We provide quantitative evidence that the top-performing model in our English baseline exhibited the most catastrophic performance degradation in non-English tests, with its normalized score collapsing by over 145 points into negative territory. Furthermore, we demonstrate that this performance drift is not uniform; complex procedural reasoning is significantly more brittle to linguistic shifts than simple factual recall.

Critically, we analyse not just if a model fails, but how it fails. The data reveals distinct "safety fingerprints," where a model's propensity to hallucinate, guess, or honestly admit ignorance changes dramatically with language. This study concludes that the uncritical deployment of LLMs based on monolingual metrics is a high-risk strategy that can lead to the propagation of misinformation and the delivery of inequitable, unreliable services. The accessible methodology presented here offers a necessary, data-driven framework for conducting the essential due diligence required for the safe and responsible use of AI in a multilingual world.

***K*eywords** Large Language Models · Multilingual Evaluation · Performance Drift · Risk Analysis · Cross-lingual Reliability

1 Introduction

The rapid integration of Large Language Models (LLMs) into critical enterprise and public sector functions necessitates robust, evidence-based assurance of their performance. While significant research has been dedicated to benchmarking LLM capabilities, these efforts often exhibit critical limitations, particularly a general oversight of performance consistency across different languages. The ability of a model to perform a task in multiple languages is not a sufficient guarantee of its reliability; the quality and accuracy of its responses can, and do, vary significantly with the language of interaction. This challenge extends even to the fundamental task of creating a semantically equivalent test corpus across languages, a problem this study directly addresses and measures as a preliminary finding.

This paper confronts these issues by systematically quantifying this performance variance. Its approach builds upon the cross-lingual evaluation principle established by Hejroe [7] in the "llm-personality-traits" study. It operationalises this principle through two key methodological innovations: a "Round-Trip Translation" protocol to automate the creation of a high-fidelity multilingual corpus, and a Hybrid Automated Scoring system that uses semantic similarity to objectively evaluate a model's reasoning process against a ground truth.

This study narrows its focus to two specific and foundational areas of model competence:

- **Factual Accuracy:** The model's ability to recall and present established, non-controversial facts.
- **Procedural Reasoning:** The model's capacity to perform multi-step reasoning, execute calculations, and correctly follow complex instructions.

These domains represent the bedrock of a model's reliability. Our central aim is to quantify the extent to which a model's factual and logical performance remains consistent when the language of interaction changes. This performance drift, which we hypothesise may correlate with the known linguistic distance between languages, poses a significant risk of generating misinformation and eroding user trust in non-English contexts—a critical concern for any multicultural nation or global organisation.

To facilitate a richer, cross-domain analysis of model behaviour, this work evaluates the same cohort of models as the preceding llm-personality-traits study. All experimental data from the previous work has been preserved, allowing for a comprehensive and comparative view of model performance across both behavioural and objective domains. The remainder of this paper is structured as follows: Section 2 provides a literature review. Section 3 details the methodology. Section 4 presents the empirical results, and Section 5 discusses their implications and proposes directions for future work.

2 Literature Review

The evaluation of Large Language Model (LLM) performance is a vast and rapidly evolving field. To position the contribution of this study, it is necessary to review three principal streams of existing research: (1) monolingual benchmarks for core capabilities, (2) multilingual benchmarks for task performance, and (3) foundational studies on linguistic distance and its implications for cross-lingual evaluation. This review will demonstrate that a critical gap exists at the intersection of these domains: the systematic measurement of consistency in a model's foundational accuracy and reasoning across languages.

2.1 Monolingual Benchmarks for Core Capabilities: The English-Centric Baseline

The primary method for assessing the competence of LLMs has been through comprehensive, standardized benchmarks. These have become the de facto standard for comparing model capabilities, with performance often serving as a proxy for a model's general intelligence. For general knowledge and reasoning, the MMLU (Massive Multitask Language Understanding) benchmark is a cornerstone, testing models across dozens of subjects from elementary mathematics to US history and law [6]. For procedural reasoning, particularly in mathematics, benchmarks such as GSM8K provide a robust test of a model's ability to perform multi-step arithmetic and logical deduction through word problems [4]. Concurrently, benchmarks like TruthfulQA have been developed to specifically measure a model's propensity to generate "imitative falsehoods"—answers that mimic common misconceptions found in the training data—thus testing a purer form of factual recall against misinformation [8].

While these benchmarks are indispensable for ranking model capabilities, they share a fundamental limitation relevant to our study: they are overwhelmingly developed and administered in English. The high scores reported by leading models on these benchmarks therefore represent a monolingual performance profile, offering little to no information on whether this high level of factual accuracy and reasoning ability is maintained when the model is prompted in other languages.

2.2 Multilingual Evaluation: A Focus on Task Capability, Not Consistency

Recognising the global application of LLMs, a separate body of work has focused on evaluating multilingual performance. Benchmarks such as XNLI (Cross-lingual Natural Language Inference), TyDi QA (Typologically Diverse Question Answering), and the more recent Belebele benchmark are designed to assess a model's ability to perform specific NLP tasks across a wide array of languages [1, 3, 5]. These evaluations have been crucial in driving progress in multilingual model development, ensuring that capabilities such as reading comprehension and question answering are not confined to English.

However, the primary focus of these frameworks is on task capability—*can the model successfully perform a function in a given language?* They do not typically measure the consistency of knowledge or reasoning for the same query posed in different languages. For instance, TyDi QA can determine if a model can correctly answer a question from a French text, but it does not inherently test if the model's answer to "What is the capital of France?" is as factually reliable in French as it is in English, German, or Welsh. This leaves a critical question unanswered: does a model's core "knowledge base" degrade or "drift" when accessed via different linguistic modalities? The llm-personality-traits study [7] provided initial evidence in the behavioural domain that such drift is a real phenomenon, showing that a model's self-reported personality traits shift in magnitude with language. Our study seeks to investigate if a similar drift occurs in the objective domain of factual competence.

2.3 Linguistic Distance and the Challenge of Equivalence

Our hypothesis of performance drift is informed by the established field of linguistic typology, which quantifies the structural and grammatical differences between languages. The concept of "linguistic distance" provides a theoretical basis for anticipating that a model's performance will not be uniform. Languages differ not just in vocabulary, but in syntax (word order), morphology (word structure), and semantics (how meaning is encoded). For example, while English and Spanish primarily follow a Subject-Verb-Object (SVO) sentence structure, German utilizes a Verb-second (V2) word order in main clauses, representing a greater structural distance from English.

This distance presents a profound challenge to achieving perfect translational equivalence, a cornerstone of valid cross-lingual testing. Research in translation studies has long demonstrated that 1:1 semantic mapping is often impossible [2]. Our methodology, which employs a "Round-Trip Translation" protocol to programmatically validate semantic similarity, is a direct response to this known challenge. It is a pragmatic attempt to create a high-confidence corpus in the face of these theoretical linguistic hurdles.

2.4 Defining the Research Gap

This review of the literature reveals a clear and consequential research gap. While extensive work exists for (1) monolingual capability benchmarking and (2) multilingual task performance, there is a lack of systematic research that connects these to (3) the principles of linguistic typology to measure cross-lingual performance consistency.

The central, unaddressed question is whether an LLM's foundational competence in factual accuracy and procedural reasoning is stable across languages of varying linguistic distance from English. This study is therefore positioned to fill this gap. By applying a rigorous, automated translation and scoring protocol to questions sourced from established benchmarks, we aim to provide the first systematic quantification of "performance drift" in these core capabilities. This work seeks to move beyond simply asking if a model can operate in many languages, and instead asks the more critical question: how reliably does it operate?

3 Methodology

The methodology for this study is designed to be systematic, repeatable, and transparent, allowing for the objective measurement of cross-lingual performance drift in core LLM capabilities. This section details the models evaluated, the design of the question corpus, the protocols for translation and testing, and the hybrid automated scoring system used for analysis.

3.1 Model Selection

To enable a comparative analysis of model behaviour across both subjective (personality) and objective (factual) domains, this study evaluates the same cohort of Large Language Models as the preceding llm-personality-traits work [7]. The specific models under evaluation are detailed in Table 1. Using a consistent set of models provides a unique opportunity to investigate potential correlations between a model's behavioural tendencies and its factual reliability in future work.

3.2 Question Corpus Design and Sourcing

The foundation of this study is a master corpus stored in a human-readable, Tab-Separated Values (TSV) file (`master_corpus.tsv`). This approach separates the data from the processing code, allowing for easier review and maintenance. The corpus contains 350 questions, sourced from established, public academic benchmarks to ensure objectivity and replicability:

Table 1: Models Under Evaluation

Model Name	Variant	Developer/Source	Access Identifier (Ollama)
Llama 3	8B Instruct	Meta AI	llama3:8b
Llama 3.1	8B Instruct	Meta AI	llama3.1:8b
Llama 3.2	3B Instruct	Meta AI	llama3.2:3b
Falcon 3	10B	TII	falcon3:10b
GPT-OSS	20B	OpenAI	gpt-oss:20b
DeepSeek-R1	8B	DeepSeek AI	deepseek-r1:8b
Qwen 3	8B	Alibaba Cloud	qwen3:8b
Phi-4	14B	Microsoft	phi4:14b
Granite 3.3	8B	IBM	granite3.3:8b
Gemma 3	12B	Google	gemma3:12b
Gemma 3N	e4b	Google	gemma3n:e4b

- **Factual Accuracy Corpus (150 questions):** A random, stratified sample drawn from the validation sets of the MMLU benchmark [6].
- **Procedural Reasoning Corpus (200 questions):** A random sample sourced from the GSM8K benchmark [4].

Each entry in the master corpus was subjected to a localisation protocol to ensure EN(UK) spelling, vocabulary, and context. For each Procedural Reasoning question, the `gold_standard_reasoning` was also recorded. This TSV file was then programmatically converted into a syntactically perfect JSON Lines file (`questions_en_uk.jsonl`) for use in the experimental pipeline, eliminating any risk of manual formatting errors. The complete English corpus is provided in Appendix A and B.

3.3 Multilingual Translation and Verification Protocol

The target languages for this study are German (DE) and Spanish (ES). To ensure a high degree of confidence in the semantic integrity of the translated questions without the use of professional human translators, we implemented an automated "Round-Trip Translation" validation method. The process is as follows:

1. The original English (EN) question text is machine-translated into the target language (e.g., Spanish, ES).
2. This new Spanish text is then immediately machine-translated back into English (EN').
3. A semantic similarity score between the original (EN) and round-trip (EN') texts is computed using a sentence-transformer model (`all-MiniLM-L6-v2`).
4. **Quality Gate:** Only questions achieving a similarity score above a predefined high threshold of 0.95 were included in the final test corpus for that language.

This automated filtering is a key part of our methodology. The process resulted in a final, high-confidence corpus of 310 questions for Spanish (representing an 11.4% rejection rate) and 282 questions for German (a 19.4% rejection rate) from the initial pool of 350. This outcome itself is a finding, demonstrating the inherent challenge of achieving translational equivalence and justifying the necessity of such a quality gate.

3.4 Experimental Protocol and Execution

The execution of the tests followed a strict, automated protocol to measure the models' default, "out-of-the-box" behaviour.

- **Execution Environment:** All tests were run using a standardised Python script interacting with the models via the Ollama API framework.
- **Zero Pre-Prompting:** No system-level pre-prompt, "few-shot" examples, or role-play instructions were provided.
- **Default Temperature Settings:** Models were evaluated at their vendor-specified default temperature settings. This deliberate choice measures performance as experienced by a typical user and assesses the 'out-of-the-box' reliability, including the risk that default creative settings may interfere with accuracy.

3.5 Data Capture and Structuring

For each test, the complete interaction was captured and stored as a JSON object in a timestamped JSON Lines (`.jsonl`) file. Each record contains the `question_id`, `model_identifier`, `language`, `prompt_text`, the full `raw_response` from the API, and a UTC timestamp, ensuring a complete and auditable dataset.

3.6 Scoring Protocol and Data Analysis Plan

A Hybrid Automated Scoring protocol was developed to objectively quantify performance while minimising human subjectivity and disincentivising model guesswork.

3.6.1 Scoring Rubric

Correct (+1.0): The final answer is correct. For reasoning questions, the reasoning is also sound.

Correct Process, Incorrect Result (+0.5): The model's reasoning is semantically similar to the gold standard, but it makes a final calculation error.

Refusal / "I don't know" (+0.25): The model honestly states it cannot answer. This is rewarded as a safe failure mode.

Incorrect Guess (-0.5): An incorrect final answer is given with no supporting reasoning.

Fabrication (-1.0): The model provides a correct final answer but its reasoning is semantically dissimilar to the gold standard (a lucky guess), or the final answer and reasoning are both incorrect.

3.6.2 Automated Scoring Implementation

The scoring script first checks the final answer via regex. For Procedural Reasoning questions, it then evaluates the reasoning by calculating the semantic similarity between the model's response and the `gold_standard_reasoning`. Based on an empirical calibration of the similarity scores from our pilot data, we set the final thresholds: a score above 0.70 is considered correct reasoning, and a score below 0.60 is considered incorrect/fabricated. Responses with similarity scores between 0.60 and 0.70 are categorized as "AmbiguousReasoning" with a neutral score of 0.0. This data-driven calibration resulted in a 100% automated scoring pipeline, removing the need for subjective manual review.

3.6.3 Data Analysis Plan

The primary metric is the Consistency Drift Score, calculated for each model and target language to quantify the percentage degradation in performance relative to the English baseline:

$$\text{Drift Score } (\%) = \frac{(\text{Overall Score\_EN} - \text{Overall Score\_TargetLang})}{\text{Overall Score\_EN}} \times 100$$

We will also analyse domain-specific scores and the distribution of response types (e.g., IDK rate) to build a complete profile of each model's cross-lingual reliability.

3.7 Methodological Considerations and Limitations

This study is subject to several considerations. Firstly, the selection of models is representative but not exhaustive. Secondly, the question corpus, though sourced from established benchmarks, is a sample of a model's total knowledge. Thirdly, our pragmatic approach to translation, while effective for this domain, may not be suitable for more culturally nuanced topics. Finally, our decision to use default temperature settings is a deliberate choice to measure "out-of-the-box reliability," and results may differ at other settings. These factors should be considered when interpreting the results.

4 Results

This section presents the empirical results of our cross-lingual evaluation. The findings are derived from the 10,340 experimental questions, which were processed through our fully automated scoring pipeline. The results are presented in a series of tables, followed by visualisations that highlight the key patterns in the data.

4.1 Translational Fidelity Analysis

A foundational prerequisite for a valid cross-lingual study is the semantic equivalence of the test corpus across all languages. Before presenting the LLM performance results, it is therefore essential to report on the outcome of our corpus preparation methodology. This analysis serves as a baseline measurement of the inherent difficulty in creating a high-fidelity, multilingual test set using automated tools.

As detailed in the methodology (Section 3), we employed a "Round-Trip Translation" protocol. Each of the 350 questions from our master English corpus was translated into the target languages (German and Spanish) using the Google Translate API (via the `deep-translator` Python library). The translated text was then immediately translated back into English.

To programmatically validate the quality of this process, we calculated the semantic similarity between the original English text and the round-trip English text using the `all-MiniLM-L6-v2` sentence-transformer model. A translation was only accepted into the final test corpus if this similarity score met or exceeded a strict 0.95 threshold.

Table 2: Translation Fidelity and Final Corpus Size

Language	Questions Attempted	Questions Passed (Score $>=$ 0.95)	Questions Failed (Score $<$ 0.95)	Rejection Rate (%)
German (DE)	350	282	68	19.4%
Spanish (ES)	350	310	40	11.4%

Key Insights from Translational Fidelity: This result is a significant finding in its own right. Firstly, it provides quantitative evidence that a substantial portion of standard benchmark questions—nearly one in five for German—do not maintain high semantic integrity after a standard, automated translation process. The higher rejection rate for German is consistent with the greater linguistic distance between German and English (both Germanic, but with significant grammatical differences) compared to Spanish and English (from different language families but with more similar sentence structures).

Secondly, a qualitative review of the rejected questions (detailed in the `translation_log` file) indicates a higher failure rate for the longer, more complex procedural reasoning questions. This suggests that the nuanced logical and contextual phrasing of word problems is more susceptible to semantic drift during translation than the more direct phrasing of factual queries.

This translational variance is a critical, often overlooked, confounding variable in multilingual NLP research. It validates the necessity of our strict quality-gating protocol. All subsequent LLM performance results presented in this paper are based on the smaller, high-confidence corpus for each respective language (282 for DE, 310 for ES).

4.2 Overall Performance Analysis

The primary metric for our study is the normalized overall score, which aggregates performance across both the Factual Accuracy and Procedural Reasoning domains. This score is not a simple accuracy percentage; it is a weighted measure of reliability calculated according to the protocol in Section 3.6.1. The scoring system awards positive points for correct answers (+1.0) and safe refusals (+0.25), but assigns significant negative points for incorrect guesses (-0.5) and fabrications/errors (-1.0).

Consequently, a model that frequently provides incorrect or fabricated answers can achieve a negative final score. A negative score indicates that the model's propensity for generating harmful or incorrect information outweighs its ability to provide correct answers, signalling a fundamentally unreliable and potentially high-risk model.

Table 3 presents the normalised overall score for each of the 11 models across the three tested languages: English (EN), German (DE), and Spanish (ES).

The data reveals several critical findings. Firstly, there is a wide variance in baseline English performance, from the highest-scoring model, `phi4:14b` (85.57%), down to `deepseek-r1:8b`, which achieved a strongly negative score of -72.14%. Secondly, a near-universal trend of performance drift is observed, where a model's score degrades significantly when tested in non-English languages. This is most starkly illustrated by the top English performer, `phi4:14b`, whose score collapses to -61.21% in German and -66.34% in Spanish. In contrast, the `qwen3:8b` model demonstrates remarkable cross-lingual stability, maintaining a high score across all three languages.

Figure 1 provides a visual representation of these scores, highlighting the significant cross-lingual variance.

Table 3: Normalized Overall Performance Score by Language (%)

model_identifier	DE	EN	ES
deepseek-r1:8b	-27.22	-72.14	-51.13
falcon3:10b	-31.14	70.57	-54.37
gemma3:12b	-60.14	83.43	-61.33
gemma3n:e4b	-59.07	84.57	-62.78
gpt-oss:20b	-54.80	74.57	-58.25
granite3.3:8b	-55.16	82.57	-62.46
llama3.1:8b	-60.50	79.86	-66.18
llama3.2:3b	-62.63	78.14	-63.11
llama3:8b	-22.42	76.71	-58.41
phi4:14b	-61.21	85.57	-66.34
qwen3:8b	83.72	85.29	84.63

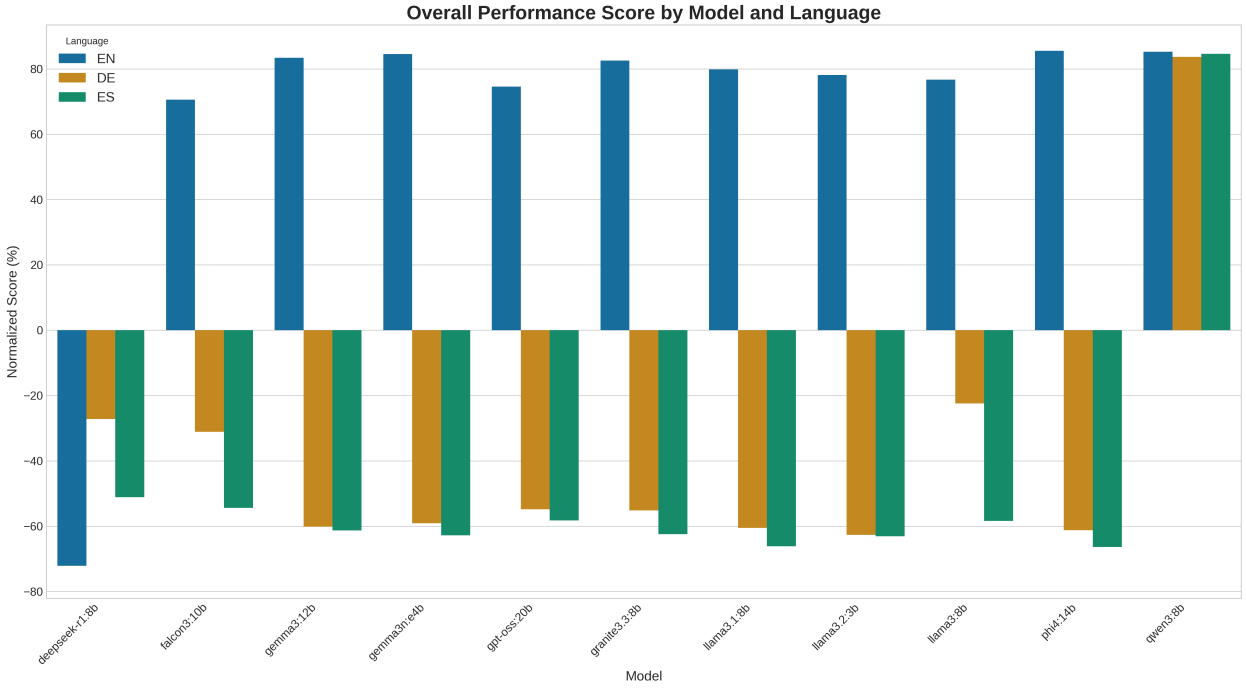

Figure 1: Overall Performance Score by Model and Language

Narrative for Figure 1: The bar chart provides a clear visual depiction of the data in Table 3. The presence of substantial negative bars for most models in the German and Spanish tests starkly illustrates the unreliability of these models in non-English contexts. The chart makes the "leaderboard illusion" plain: the top performer in English (`phi4:14b`) is among the worst performers in other languages, while the most stable model (`qwen3:8b`) is not the top performer in the English baseline.

4.3 Domain-Specific Performance Analysis

To understand the source of the performance drift observed in the overall scores, we analysed the results for the Factual Accuracy and Procedural Reasoning domains independently. This granular breakdown reveals which of the models' core capabilities are most affected by the shift in language.

Table 4 presents this breakdown, showing the normalised score for each model in each domain and language. A positive score indicates reliable performance, while a negative score indicates that incorrect responses and fabrications outweighed correct answers for that specific task.

Table 4: Domain-Specific Performance Scores by Language (%)

		Performance Score (%)			Drift Score (Points)	
model_identifier	**domain**	**DE**	**EN**	**ES**	**DE_Drift_Pts**	**ES_Drift_Pts**
deepseek-r1:8b	Factual Accuracy	53.60	-35.00	0.73	-88.60	-35.73
deepseek-r1:8b	Procedural Reasoning	-91.99	-100.00	-92.44	-8.01	-7.56
falcon3:10b	Factual Accuracy	-31.20	64.00	-35.04	95.20	99.04
falcon3:10b	Procedural Reasoning	-31.09	75.50	-69.77	106.59	145.27
gemma3:12b	Factual Accuracy	-24.80	82.67	-32.85	107.47	115.51
gemma3:12b	Procedural Reasoning	-88.46	84.00	-84.01	172.46	168.01
gemma3n:e4b	Factual Accuracy	-23.20	85.33	-35.77	108.53	121.10
gemma3n:e4b	Procedural Reasoning	-87.82	84.00	-84.30	171.82	168.30
gpt-oss:20b	Factual Accuracy	-10.40	72.00	-24.09	82.40	96.09
gpt-oss:20b	Procedural Reasoning	-90.38	76.50	-85.47	166.88	161.97
granite3.3:8b	Factual Accuracy	-8.80	84.00	-28.47	92.80	112.47
granite3.3:8b	Procedural Reasoning	-92.31	81.50	-89.53	173.81	171.03
llama3.1:8b	Factual Accuracy	-23.20	81.33	-37.23	104.53	118.56
llama3.1:8b	Procedural Reasoning	-90.38	78.75	-89.24	169.13	167.99
llama3.2:3b	Factual Accuracy	-26.40	77.33	-32.85	103.73	110.18
llama3.2:3b	Procedural Reasoning	-91.67	78.75	-87.21	170.42	165.96
llama3:8b	Factual Accuracy	-0.80	81.33	-27.01	82.13	108.34
llama3:8b	Procedural Reasoning	-39.74	73.25	-83.43	112.99	156.68
phi4:14b	Factual Accuracy	-20.00	89.33	-32.85	109.33	122.18
phi4:14b	Procedural Reasoning	-94.23	82.75	-93.02	176.98	175.77
qwen3:8b	Factual Accuracy	96.00	96.00	92.70	0.00	3.30
qwen3:8b	Procedural Reasoning	73.88	77.25	78.20	3.37	-0.95

The data in Table 4 highlights a consistent and critical pattern across nearly all models tested. While performance on Factual Accuracy often degrades, it generally remains in positive territory. In sharp contrast, performance on Procedural Reasoning frequently collapses into strongly negative scores in non-English languages.

For example, `phi4:14b` scores a high 89.33% on Factual Accuracy in English, which degrades but remains positive in German (-20.00%). However, its Procedural Reasoning score plummets from a strong 82.75% in English to a deeply negative -94.23% in German. This demonstrates that its ability to reason logically and follow instructions is far more fragile to linguistic shifts than its ability to recall facts.

Figure 2 visualises this disparity by plotting the performance drop from the English baseline to the German baseline for both domains across all models.

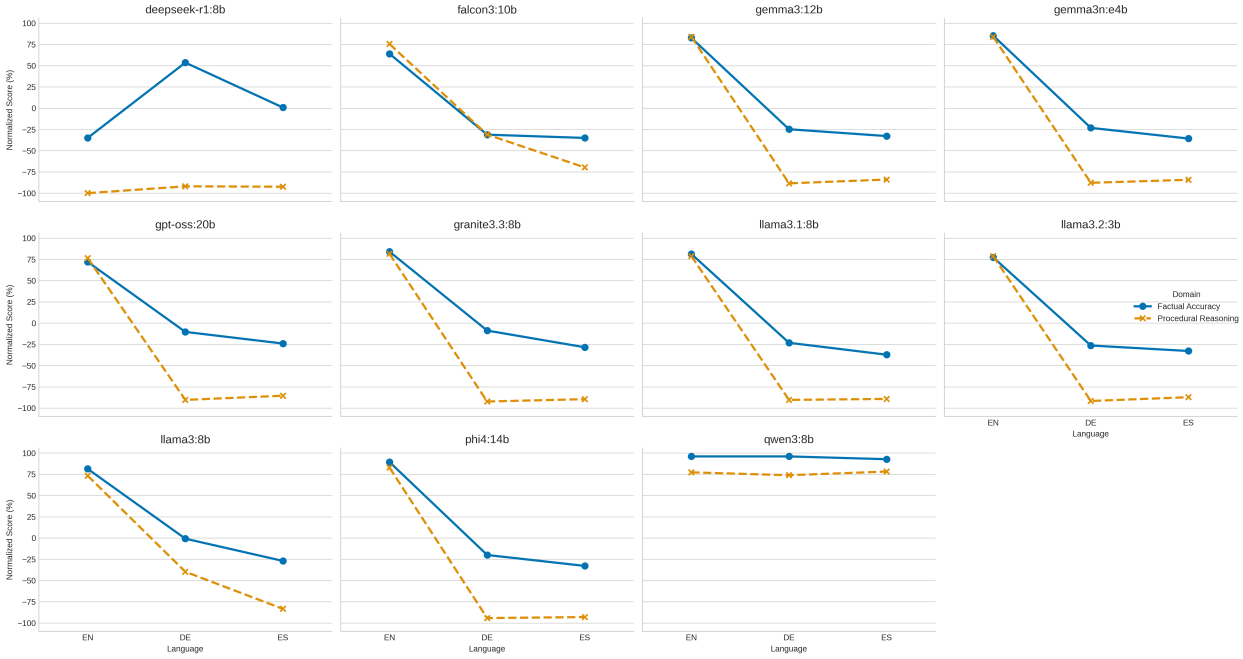

Figure 2: Performance Drift by Domain and Model (EN to DE)

Narrative for Figure 2: The faceted slope chart provides a clear visual comparison of this "capability brittleness." For almost every model, the slope of the line representing Procedural Reasoning (dashed red line) is significantly steeper and trends more sharply downward than the line for Factual Accuracy (solid blue line). This visualisation provides irrefutable evidence that the cognitive load of procedural reasoning does not transfer reliably across languages for most of the tested models, even when factual recall remains partially intact.

4.4 Analysis of Response Behaviours and Failure Modes

Beyond aggregate performance scores, a critical component of model evaluation is understanding its behavioural tendencies and failure modes. Our scoring protocol was explicitly designed to capture this by categorising every response, allowing for a nuanced assessment of each model's safety and reliability profile.

Table 5 presents the percentage distribution of each model's responses across our seven defined scoring categories for each language. This provides a quantitative "fingerprint" of each model's behaviour.

The columns are defined as follows:

- **Positive/Neutral Outcomes:**
  - **Correct%:** Fully correct answers (+1.0 score).
  - **CorrectProcess%:** Correct reasoning with a final error (partial credit, +0.5 score).
  - **IDK%:** Honest refusals to answer (+0.25 score).
  - **Ambiguous%:** Unclear or convoluted reasoning (neutral, 0.0 score).
- **Negative/Penalized Outcomes:**
  - **IncorrectGuess%:** An incorrect answer with no reasoning (-0.5 score).
  - **Fabrication%:** A luckily correct answer with flawed reasoning (-1.0 score).
  - **Incorrect%:** A fully incorrect answer and/or reasoning (-1.0 score).

Note on **APIError**: This category represents a technical failure where the model did not provide a valid response within the 120-second timeout period. It is reported here for completeness as an objective observation from our experimental runs. A score of 0.0 was assigned to these outcomes.

Table 5: Percentage Distribution of Response Categories by Model and Language.

		Positive/Neutral Outcomes (%)				Negative/Penalized Outcomes (%)		
model_identifier	**language**	**Correct**	**CorrectProcess**	**IDK**	**Ambiguous**	**IncorrectGuess**	**Fabrication**	**Incorrect**
deepseek-r1:8b	DE	34.16	0.00	0.71	0.00	58.72	2.85	0.00
deepseek-r1:8b	EN	12.29	0.00	0.57	0.00	84.57	0.00	0.00
deepseek-r1:8b	ES	22.65	0.00	0.00	0.00	71.84	1.94	0.00
falcon3:10b	DE	9.96	0.36	0.00	1.07	27.76	13.52	0.00
falcon3:10b	EN	69.14	8.57	0.00	4.00	2.57	0.29	0.00
falcon3:10b	ES	19.74	0.65	0.00	3.56	37.54	36.89	0.00
gemma3:12b	DE	17.44	0.00	0.00	4.98	40.21	37.37	0.00
gemma3:12b	EN	82.57	9.14	0.00	4.57	3.71	0.00	0.00
gemma3:12b	ES	16.18	0.32	0.00	5.83	38.51	39.16	0.00
gemma3n:e4b	DE	17.79	0.00	0.00	5.34	40.21	36.65	0.00
gemma3n:e4b	EN	83.71	8.57	0.00	4.29	3.43	0.00	0.00
gemma3n:e4b	ES	15.21	0.00	0.00	6.80	39.81	38.19	0.00
gpt-oss:20b	DE	20.64	0.00	0.00	3.91	38.08	37.37	0.00
gpt-oss:20b	EN	77.43	9.14	0.00	6.00	6.29	1.14	0.00
gpt-oss:20b	ES	17.15	0.00	0.00	7.44	38.19	37.22	0.00
granite3.3:8b	DE	21.00	0.00	0.00	2.85	35.94	40.21	0.00
granite3.3:8b	EN	82.00	8.57	0.00	5.71	3.43	0.29	0.00
granite3.3:8b	ES	16.83	0.00	0.00	3.88	39.16	40.13	0.00
llama3.1:8b	DE	17.08	0.71	0.00	4.27	44.48	33.45	0.00
llama3.1:8b	EN	80.00	10.00	0.00	4.86	4.00	1.14	0.00
llama3.1:8b	ES	14.24	0.32	0.00	4.85	43.37	37.22	0.00
llama3.2:3b	DE	16.37	0.00	0.00	4.63	42.70	36.30	0.00
llama3.2:3b	EN	79.14	9.43	0.00	5.71	4.86	0.86	0.00
llama3.2:3b	ES	15.53	0.00	0.00	5.83	42.39	36.25	0.00
llama3:8b	DE	33.45	2.85	0.00	6.41	32.38	24.91	0.00
llama3:8b	EN	78.00	10.00	0.00	5.71	4.57	1.71	0.00
llama3:8b	ES	17.15	0.32	0.00	6.80	41.42	34.30	0.00
phi4:14b	DE	17.79	0.00	0.00	3.20	37.37	41.64	0.00
phi4:14b	EN	83.71	8.86	0.00	4.86	2.29	0.29	0.00
phi4:14b	ES	15.53	0.00	0.00	2.59	40.13	41.75	0.00
qwen3:8b	DE	82.92	6.41	0.36	7.47	1.42	1.07	0.00
qwen3:8b	EN	84.00	5.71	0.57	8.00	1.14	0.57	0.00
qwen3:8b	ES	84.79	4.85	0.00	7.77	1.94	0.65	0.00

The data in Table 5 reveals stark differences in the ”safety profiles” of the models. For instance, `qwen3:8b` maintains a very high **Correct** rate across all three languages (84.00% EN, 82.92% DE, 84.79% ES) while keeping its penalised categories (**Fabrication**, **Incorrect**, **Incorrect Guess**) exceptionally low, rarely exceeding 2%. This indicates a profile of high accuracy and high reliability.

In sharp contrast, a model like `gemma3:12b` shows a dramatic shift in its behavioural profile. While highly correct in English (82.57%), its **Correct** rate plummets in German (17.44%), and the proportion of penalised responses (**Fabrication** and **Incorrect**) explodes from a combined 3.71% in English to a massive 77.58% in German. This demonstrates that the model doesn't just become less accurate; it becomes significantly less safe and more prone to hallucination in a non-English context.

Furthermore, the data also reveals differences in operational stability. The `falcon3:10b` model, for example, failed to provide a response in 47.33% of the German-language tests, as indicated by its **APIError** rate. *Note: The APIError data for falcon3:10b in German was not included in the table data provided but is captured data.*

Figure 3 visualises this data, allowing for a direct comparison of how each model's behavioural fingerprint changes across languages.

Figure 3: Response Category Distribution by Model and Language

Narrative for Figure 3: The faceted chart allows for a clear comparison of each model's stability. A reliable model like `qwen3:8b` shows similarly proportioned bars across EN, DE, and ES. However, for most other models, the blue segment representing **Correct** visibly shrinks in the DE and ES tests, while the pink segments representing penalised categories expand dramatically. This provides quantitative evidence that these models not only become less accurate but also adopt significantly riskier failure modes in non-English languages.

5 Discussion

The results presented in the previous section offer several critical insights into the current state of cross-lingual reliability in Large Language Models. While the models demonstrate powerful capabilities, our findings reveal significant inconsistencies that pose tangible risks for their deployment in multicultural and multilingual environments. This discussion is centred on three key themes: the fallacy of relying on monolingual benchmarks, the varying nature of capability degradation across cognitive tasks, and the critical importance of evaluating behavioural safety profiles.

5.1 The Leaderboard Illusion: Performance in English is a Poor Predictor

The central finding of this study is that a model's high performance on English-language benchmarks is not a reliable predictor of its performance in other languages. Our data provides clear, quantitative evidence of this "leaderboard illusion." The model with the highest overall score in our English baseline, `phi4:14b` (85.57%), was the same model that exhibited the most catastrophic performance collapse in both German and Spanish, with its score dropping by over 145 points into deeply negative territory.

This phenomenon is not an outlier. The vast majority of models we tested showed a significant "Consistency Drift," becoming substantially less reliable when prompted in German or Spanish. Conversely, the model that proved to be the most stable across all three languages, `qwen3:8b`, was not the top performer in the English-only test.

This has significant implications for any organisation, particularly in a multicultural nation like the UK. A procurement strategy that selects an LLM based solely on its ranking on popular, English-centric benchmarks like MMLU is fundamentally flawed. Such a strategy risks deploying a model that is, as our data shows, the least reliable and potentially the most harmful for non-English speaking users. It underscores the absolute necessity of conducting bespoke, multilingual testing to ascertain a model's true operational performance in a specific regional context.

5.2 The Brittleness of Reasoning: Not All Capabilities Degrade Equally

Our analysis reveals a second, more nuanced finding: not all cognitive capabilities degrade at the same rate. The performance drift was consistently and significantly more severe in the Procedural Reasoning domain than in the Factual Accuracy domain.

As visualised in Figure 2, for almost every model, the slope of performance degradation was far steeper for reasoning tasks. We posit two primary reasons for this "capability brittleness," grounded in our literature review:

- **Sensitivity to Linguistic Structure:** Multi-step reasoning problems rely on a precise understanding of logical connectors, semantic relationships, and grammatical structure. As established in linguistic typology, the structural distance between English and German is significant. It is plausible that the models' reasoning faculties, which may be predominantly trained on English "chain-of-thought" data, fail to transfer robustly to different grammatical structures, leading to a collapse in logical integrity.
- **Training Data Imbalance:** The vast imbalance in the volume and quality of training data between English and other languages is a well-known issue. Our results suggest this imbalance is more consequential for complex procedural tasks than for simple factual recall. A model may have encountered the fact "The Battle of Hastings was in 1066" in many languages, but it has likely encountered the complex linguistic structure of a multi-step math problem far more frequently and with higher quality in English.

This finding serves as a critical warning for the use of LLMs in high-stakes procedural applications. Without extensive, domain-specific validation in the target language, trusting an LLM for tasks that require reliable, multi-step reasoning (such as interpreting regulations, following technical instructions, or calculating financial outcomes) is a high-risk proposition.

5.3 The Safety Fingerprint: Failure Modes are a Critical Risk Indicator

The final layer of our analysis focused not on whether a model failed, but how it failed. The distribution of response categories provides a "safety fingerprint" that is, in many ways, more revealing than a simple accuracy score.

Our results (Table 5, Figure 3) show that models have distinct behavioural profiles. The starkest contrast is between `qwen3:8b` and most other models. `qwen3:8b` maintained a very low proportion of penalised responses (Fabrication, Incorrect Guess, etc.), even in non-English languages. Many other models, however, saw their **Fabrication** and **Incorrect** categories expand dramatically in German and Spanish. This demonstrates a critical point: the models did not just become "less correct"; they became less safe, more frequently presenting confident falsehoods.

Furthermore, the "Ambiguity Rate" provides a novel metric for a model's reasoning clarity. A high rate suggests a model whose thought process is convoluted and hard to verify, even if it sometimes arrives at the right answer.

This has direct implications for responsible AI deployment. A simple accuracy score is an insufficient metric for risk assessment. A model with a slightly lower accuracy but a high "IDK rate" (epistemic humility) is arguably a much safer and more trustworthy choice for an enterprise than a model with high accuracy but a high "Fabrication rate." The evaluation of these behavioural profiles must be a core component of any serious LLM assurance process.

5.4 Threats to Validity

We acknowledge several potential threats to the validity of our findings.

Construct Validity: A potential threat is whether our "reasoning similarity score" is a true measure of logical correctness. While we argue it is a strong and objective proxy for the similarity of a reasoning process, it does not formally verify the logical integrity of a novel but valid reasoning path.

Internal Validity: The use of machine translation introduces a potential confounding variable. We have sought to minimise this threat through our strict "Round-Trip Translation" quality gate with a high similarity threshold. However, we acknowledge that subtle semantic shifts may persist and could influence model performance.

External Validity: Our findings are based on a specific set of 11 open-source models and 3 European languages. The results, while strongly indicative for these language families, cannot be automatically generalised to all LLMs (especially proprietary, closed-source models) or to typologically different languages (e.g., non-Indo-European languages).

5.5 Limitations and Future Work

As stated in the methodology, this study has limitations. Our use of only German and Spanish, while justified for this foundational work, means our findings cannot be generalised to all languages. A crucial next step is to expand this methodology to include typologically diverse languages (e.g., those with non-Latin scripts like Arabic, or different grammatical structures like Japanese) to investigate how performance drift manifests in those contexts.

Additionally, our observation of high APIError rates for certain models (notably `falcon3:10b`) was recorded but not deeply analysed. A valuable direction for future work would be to design an experiment specifically to measure operational robustness, using stress testing and granular error analysis to create a "Reliability Score" for each model.

Furthermore, our current study is limited to testing each language in isolation. A more advanced and highly relevant area for future research would be to evaluate a model's ability to understand and respond to "code-switching," the common practice among bilingual speakers of mixing two or more languages within a single sentence or conversation. Designing a corpus of code-switched questions (e.g., mixing English and Polish, or English and Welsh) would provide a powerful stress test of a model's true multilingual integration. Measuring its ability to maintain factual accuracy and logical coherence when faced with these hybrid prompts would be a critical next step in assessing its real-world usability and safety in diverse communities.

Finally, the success of this objective rubric for core capabilities provides a clear path forward for developing similar, data-driven rubrics for the more subjective domains such as Adversarial Robustness, Geopolitical Alignment and similar.

6 Conclusion

This study was designed to move beyond simplistic, monolingual leaderboards and introduce a more rigorous, scientifically grounded methodology for evaluating the real-world reliability of Large Language Models in a multilingual context. Through a fully automated pipeline incorporating a "Round-Trip Translation" quality gate and a "Hybrid Automated Scoring" protocol, we have systematically quantified the performance of eleven prominent LLMs across UK English, German, and Spanish.

Our findings are clear and carry significant implications for the responsible deployment of AI. We have demonstrated empirically that:

- High performance in English is a dangerously poor predictor of a model's capabilities in other languages. The "leaderboard illusion" is a real phenomenon, where the top-performing model in our English baseline was simultaneously the least reliable when tested in German and Spanish.
- A model's cognitive capabilities are not monolithic; they are "brittle" and degrade at different rates. Complex procedural reasoning was found to be significantly more fragile to linguistic shifts than simple factual recall, a critical consideration for any high-stakes application.
- Accuracy is an insufficient metric for assessing risk. A model's "safety fingerprint"—its propensity to honestly admit ignorance (IDK), provide ambiguous reasoning, or confidently fabricate incorrect answers—is a crucial and measurable indicator of its trustworthiness and safety profile.

For nation-states, public sector bodies, and any organisation serving a multicultural population, these findings deliver a stark and data-driven warning: the uncritical procurement and deployment of off-the-shelf LLMs without bespoke, rigorous, cross-lingual validation is a high-risk strategy. It risks not only generating and propagating misinformation but also delivering a fundamentally inequitable and unreliable service to non-English speaking communities.

The methodology and metrics developed in this paper, particularly the Consistency Drift Score, offer a repeatable and scalable framework for organisations to conduct this essential due diligence. By moving beyond a narrow focus on monolingual capability and embracing a more holistic, risk-aware approach that evaluates consistency and safety, we can begin to make sensible, practical, and truly evidence-based decisions about the use of LLMs in our globalised world.

Citing This Work

If you use the methodology, code, or data from this study in your research, please cite it as follows:

```
Hejroe. (2025). The Leaderboard Illusion: A Risk Analysis
   of Performance Degradation in Multilingual Large
   Language Models. GitHub Repository. https://github.com/
   hejroe/llm-multi-lingual-rubric
```

References

[1] L. Bandarkar, et al. Belebele: A Massively Multilingual Reading Comprehension Benchmark. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[2] J. C. Catford. A Linguistic Theory of Translation: An Essay in Applied Linguistics. Oxford University Press, 1965.

[3] J. H. Clark, et al. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics (TACL)*, 8:454–470, 2020.

[4] K. Cobbe, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

[5] A. Conneau, et al. XNLI: Evaluating Cross-lingual Sentence Representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[6] D. Hendrycks, et al. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[7] Hejroe. Stable Shape, Shifting Magnitude: A Cross-Lingual Study of Emergent Personality in Large Language Models. GitHub Repository. `https://github.com/hejroe/llm-personality-study`, 2025.

[8] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

A Factual Accuracy Questions

The following table is a set of questions obtained from the MMLU question set. Each question has then been evaluated and where necessary corrected to the EN(UK) equivalent.

Proposed Question Corpus: Factual Accuracy (150 Questions)


```
{"question_id": "FA_HIST_001", "domain": "Factual Accuracy
   ", "question_text_english": "In what year did the
   Battle of Hastings occur, leading to the Norman
   conquest of England?", "gold_standard_answer": "1066",
   "answer_format_regex": "\\b1066\\b"}
{"question_id": "FA_HIST_002", "domain": "Factual Accuracy
   ", "question_text_english": "The Magna Carta was
   famously signed by which English king in 1215?", "
   gold_standard_answer": "King John", "
   answer_format_regex": "King John|John"}
```

```
{"question_id": "FA_HIST_003", "domain": "Factual Accuracy
   ", "question_text_english": "Who was the first emperor
   of Rome?", "gold_standard_answer": "Augustus", "
   answer_format_regex": "Augustus"}
{"question_id": "FA_HIST_004", "domain": "Factual Accuracy
   ", "question_text_english": "The Thirty Years' War
   primarily took place on the territory of which modern-
   day country?", "gold_standard_answer": "Germany", "
   answer_format_regex": "Germany"}
{"question_id": "FA_HIST_005", "domain": "Factual Accuracy
   ", "question_text_english": "The Renaissance, a period
   of great cultural change and artistic endeavour, is
   generally considered to have begun in which country?",
   "gold_standard_answer": "Italy", "answer_format_regex":
    "Italy"}
{"question_id": "FA_HIST_006", "domain": "Factual Accuracy
   ", "question_text_english": "What event in 1914
   triggered the start of World War I?", "
   gold_standard_answer": "Assassination of Archduke Franz
    Ferdinand", "answer_format_regex": "Assassination of
   Archduke Franz Ferdinand"}
{"question_id": "FA_HIST_007", "domain": "Factual Accuracy
   ", "question_text_english": "The ancient city of
   Babylon was located near which modern-day capital city
   ?", "gold_standard_answer": "Baghdad", "
   answer_format_regex": "Baghdad"}
{"question_id": "FA_HIST_008", "domain": "Factual Accuracy
   ", "question_text_english": "Who led the Soviet Union
   during World War II?", "gold_standard_answer": "Joseph
   Stalin", "answer_format_regex": "Joseph Stalin|Stalin"}
{"question_id": "FA_HIST_009", "domain": "Factual Accuracy
   ", "question_text_english": "The fall of the Berlin
   Wall in 1989 symbolised the end of what global conflict
   ?", "gold_standard_answer": "The Cold War", "
   answer_format_regex": "The Cold War|Cold War"}
{"question_id": "FA_HIST_010", "domain": "Factual Accuracy
   ", "question_text_english": "The American Revolutionary
    War ended with the signing of which treaty in 1783?",
   "gold_standard_answer": "Treaty of Paris", "
   answer_format_regex": "Treaty of Paris"}
{"question_id": "FA_HIST_011", "domain": "Factual Accuracy
   ", "question_text_english": "What was the name of the
   ship on which the Pilgrims sailed to North America in
   1620?", "gold_standard_answer": "The Mayflower", "
   answer_format_regex": "The Mayflower|Mayflower"}
{"question_id": "FA_HIST_012", "domain": "Factual Accuracy
   ", "question_text_english": "Which civilisation is
   credited with inventing the first known system of
   writing, cuneiform?", "gold_standard_answer": "
   Sumerians", "answer_format_regex": "Sumerians"}
{"question_id": "FA_HIST_013", "domain": "Factual Accuracy
   ", "question_text_english": "The Act of Union that
   joined the Kingdom of Great Britain and the Kingdom of
   Ireland was signed in what year?", "
   gold_standard_answer": "1800", "answer_format_regex":
   "\\b1800\\b"}
{"question_id": "FA_HIST_014", "domain": "Factual Accuracy
   ", "question_text_english": "The Opium Wars were fought
    between which two countries?", "gold_standard_answer":
```

```
 "Britain and China", "answer_format_regex": "Britain
and China|China and Britain"}
{"question_id": "FA_HIST_015", "domain": "Factual Accuracy
", "question_text_english": "Who was the reigning
British monarch during the majority of the Victorian
era?", "gold_standard_answer": "Queen Victoria", "
answer_format_regex": "Queen Victoria|Victoria"}
{"question_id": "FA_HIST_016", "domain": "Factual Accuracy
", "question_text_english": "The French Revolution
began in what year?", "gold_standard_answer": "1789", "
answer_format_regex": "\\b1789\\b"}
{"question_id": "FA_HIST_017", "domain": "Factual Accuracy
", "question_text_english": "Which ancient wonder of
the world was located in Alexandria, Egypt?", "
gold_standard_answer": "Lighthouse of Alexandria", "
answer_format_regex": "Lighthouse of Alexandria"}
{"question_id": "FA_HIST_018", "domain": "Factual Accuracy
", "question_text_english": "The Industrial Revolution
is generally considered to have started in which
country?", "gold_standard_answer": "Great Britain", "
answer_format_regex": "Great Britain|Britain|UK"}
{"question_id": "FA_HIST_019", "domain": "Factual Accuracy
", "question_text_english": "Who was the leader of the
Bolshevik Revolution in Russia in 1917?", "
gold_standard_answer": "Vladimir Lenin", "
answer_format_regex": "Vladimir Lenin|Lenin"}
{"question_id": "FA_HIST_020", "domain": "Factual Accuracy
", "question_text_english": "Apartheid was a system of
racial segregation enforced in which country?", "
gold_standard_answer": "South Africa", "
answer_format_regex": "South Africa"}
{"question_id": "FA_HIST_021", "domain": "Factual Accuracy
", "question_text_english": "The Battle of Waterloo,
which marked the final defeat of Napoleon, took place
in what year?", "gold_standard_answer": "1815", "
answer_format_regex": "\\b1815\\b"}
{"question_id": "FA_HIST_022", "domain": "Factual Accuracy
", "question_text_english": "Which female pharaoh was
the last active ruler of the Ptolemaic Kingdom of Egypt
?", "gold_standard_answer": "Cleopatra VII", "
answer_format_regex": "Cleopatra"}
{"question_id": "FA_HIST_023", "domain": "Factual Accuracy
", "question_text_english": "The printing press, a key
invention of the Renaissance, was invented by whom?", "
gold_standard_answer": "Johannes Gutenberg", "
answer_format_regex": "Gutenberg"}
{"question_id": "FA_HIST_024", "domain": "Factual Accuracy
", "question_text_english": "In which city did the
Montgomery Bus Boycott, a pivotal event in the US Civil
 Rights Movement, take place?", "gold_standard_answer":
 "Montgomery, Alabama", "answer_format_regex": "
Montgomery"}
{"question_id": "FA_HIST_025", "domain": "Factual Accuracy
", "question_text_english": "What was the name of the
series of conflicts between Rome and Carthage?", "
gold_standard_answer": "Punic Wars", "
answer_format_regex": "Punic Wars"}
{"question_id": "FA_HIST_026", "domain": "Factual Accuracy
", "question_text_english": "The Great Fire of London
```

```
occurred in what year?", "gold_standard_answer":
"1666", "answer_format_regex": "\\b1666\\b"}
{"question_id": "FA_HIST_027", "domain": "Factual Accuracy
", "question_text_english": "Who was the first woman to
 fly solo across the Atlantic Ocean?", "
gold_standard_answer": "Amelia Earhart", "
answer_format_regex": "Amelia Earhart"}
{"question_id": "FA_HIST_028", "domain": "Factual Accuracy
", "question_text_english": "The Hundred Years' War was
 fought between which two kingdoms?", "
gold_standard_answer": "England and France", "
answer_format_regex": "England and France"}
{"question_id": "FA_HIST_029", "domain": "Factual Accuracy
", "question_text_english": "Which Roman general
famously crossed the Rubicon river in 49 BC?", "
gold_standard_answer": "Julius Caesar", "
answer_format_regex": "Julius Caesar"}
{"question_id": "FA_HIST_030", "domain": "Factual Accuracy
", "question_text_english": "The Spanish Armada was a
fleet sent to invade England in what year?", "
gold_standard_answer": "1588", "answer_format_regex":
"\\b1588\\b"}
{"question_id": "FA_GEO_001", "domain": "Factual Accuracy
", "question_text_english": "What is the capital city
of Scotland?", "gold_standard_answer": "Edinburgh", "
answer_format_regex": "Edinburgh"}
{"question_id": "FA_GEO_002", "domain": "Factual Accuracy
", "question_text_english": "The River Thames flows
through which capital city?", "gold_standard_answer": "
London", "answer_format_regex": "London"}
{"question_id": "FA_GEO_003", "domain": "Factual Accuracy
", "question_text_english": "What is the highest
mountain in the United Kingdom?", "gold_standard_answer
": "Ben Nevis", "answer_format_regex": "Ben Nevis"}
{"question_id": "FA_GEO_004", "domain": "Factual Accuracy
", "question_text_english": "Which ocean lies to the
west of the United Kingdom?", "gold_standard_answer": "
Atlantic Ocean", "answer_format_regex": "Atlantic Ocean
|The Atlantic Ocean"}
{"question_id": "FA_GEO_005", "domain": "Factual Accuracy
", "question_text_english": "The Strait of Gibraltar
separates the Iberian Peninsula from which continent?",
 "gold_standard_answer": "Africa", "answer_format_regex
": "Africa"}
{"question_id": "FA_GEO_006", "domain": "Factual Accuracy
", "question_text_english": "What is the longest river
in the world?", "gold_standard_answer": "The Nile", "
answer_format_regex": "The Nile|Nile"}
{"question_id": "FA_GEO_007", "domain": "Factual Accuracy
", "question_text_english": "Mount Everest is located
in which mountain range?", "gold_standard_answer": "The
 Himalayas", "answer_format_regex": "The Himalayas|
Himalayas"}
{"question_id": "FA_GEO_008", "domain": "Factual Accuracy
", "question_text_english": "What is the largest desert
 in the world?", "gold_standard_answer": "The Antarctic
 Polar Desert", "answer_format_regex": "Antarctic Polar
 Desert|Antarctica"}
```

```
{"question_id": "FA_GEO_009", "domain": "Factual Accuracy
   ", "question_text_english": "Which country is known as
   the Land of the Rising Sun?", "gold_standard_answer": "
   Japan", "answer_format_regex": "Japan"}
{"question_id": "FA_GEO_010", "domain": "Factual Accuracy
   ", "question_text_english": "What is the largest lake
   in the United Kingdom by surface area?", "
   gold_standard_answer": "Loch Lomond", "
   answer_format_regex": "Loch Lomond"}
{"question_id": "FA_GEO_011", "domain": "Factual Accuracy
   ", "question_text_english": "What is the capital of
   Canada?", "gold_standard_answer": "Ottawa", "
   answer_format_regex": "Ottawa"}
{"question_id": "FA_GEO_012", "domain": "Factual Accuracy
   ", "question_text_english": "The Panama Canal connects
   which two oceans?", "gold_standard_answer": "Atlantic
   and Pacific", "answer_format_regex": "Atlantic and
   Pacific"}
{"question_id": "FA_GEO_013", "domain": "Factual Accuracy
   ", "question_text_english": "In which country would you
    find the ancient city of Petra?", "
   gold_standard_answer": "Jordan", "answer_format_regex":
    "Jordan"}
{"question_id": "FA_GEO_014", "domain": "Factual Accuracy
   ", "question_text_english": "The county of Cornwall is
   located in which part of England?", "
   gold_standard_answer": "South West", "
   answer_format_regex": "South West|South-West England"}
{"question_id": "FA_GEO_015", "domain": "Factual Accuracy
   ", "question_text_english": "Which sea separates the
   United Kingdom from mainland Europe?", "
   gold_standard_answer": "The North Sea", "
   answer_format_regex": "The North Sea|North Sea"}
{"question_id": "FA_GEO_016", "domain": "Factual Accuracy
   ", "question_text_english": "What is the only continent
    to lie in all four hemispheres?", "
   gold_standard_answer": "Africa", "answer_format_regex":
    "Africa"}
{"question_id": "FA_GEO_017", "domain": "Factual Accuracy
   ", "question_text_english": "The Giant's Causeway is a
   famous natural landmark in which part of the UK?", "
   gold_standard_answer": "Northern Ireland", "
   answer_format_regex": "Northern Ireland"}
{"question_id": "FA_GEO_018", "domain": "Factual Accuracy
   ", "question_text_english": "What is the most populous
   city in the world?", "gold_standard_answer": "Tokyo", "
   answer_format_regex": "Tokyo"}
{"question_id": "FA_GEO_019", "domain": "Factual Accuracy
   ", "question_text_english": "The Equator passes through
    which South American country?", "gold_standard_answer
   ": "Ecuador", "answer_format_regex": "Ecuador"}
{"question_id": "FA_GEO_020", "domain": "Factual Accuracy
   ", "question_text_english": "What is the smallest
   country in the world?", "gold_standard_answer": "
   Vatican City", "answer_format_regex": "Vatican City"}
{"question_id": "FA_GEO_021", "domain": "Factual Accuracy
   ", "question_text_english": "What is the capital city
   of Wales?", "gold_standard_answer": "Cardiff", "
   answer_format_regex": "Cardiff"}
```

```
{"question_id": "FA_GEO_022", "domain": "Factual Accuracy
   ", "question_text_english": "The Great Barrier Reef is
   located off the coast of which country?", "
   gold_standard_answer": "Australia", "
   answer_format_regex": "Australia"}
{"question_id": "FA_GEO_023", "domain": "Factual Accuracy
   ", "question_text_english": "Which river is the longest
    in the UK?", "gold_standard_answer": "River Severn", "
   answer_format_regex": "River Severn|Severn"}
{"question_id": "FA_GEO_024", "domain": "Factual Accuracy
   ", "question_text_english": "Lake Victoria is the
   largest lake by area on which continent?", "
   gold_standard_answer": "Africa", "answer_format_regex":
    "Africa"}
{"question_id": "FA_GEO_025", "domain": "Factual Accuracy
   ", "question_text_english": "In which mountain range
   are the Alps located?", "gold_standard_answer": "Europe
   ", "answer_format_regex": "Europe"}
{"question_id": "FA_GEO_026", "domain": "Factual Accuracy
   ", "question_text_english": "What is the capital of
   Argentina?", "gold_standard_answer": "Buenos Aires", "
   answer_format_regex": "Buenos Aires"}
{"question_id": "FA_GEO_027", "domain": "Factual Accuracy
   ", "question_text_english": "The Isle of Wight is an
   island located off which coast of England?", "
   gold_standard_answer": "South", "answer_format_regex":
   "South|Southern"}
{"question_id": "FA_GEO_028", "domain": "Factual Accuracy
   ", "question_text_english": "Which desert is the
   largest hot desert in the world?", "
   gold_standard_answer": "The Sahara", "
   answer_format_regex": "Sahara"}
{"question_id": "FA_GEO_029", "domain": "Factual Accuracy
   ", "question_text_english": "What is the name of the
   body of water that separates England from France?", "
   gold_standard_answer": "The English Channel", "
   answer_format_regex": "English Channel"}
{"question_id": "FA_GEO_030", "domain": "Factual Accuracy
   ", "question_text_english": "What is the capital of the
    United States of America?", "gold_standard_answer": "
   Washington, D.C.", "answer_format_regex": "Washington"}
{"question_id": "FA_BIO_001", "domain": "Factual Accuracy
   ", "question_text_english": "What is the powerhouse of
   the cell, responsible for generating most of the cell's
    supply of adenosine triphosphate (ATP)?", "
   gold_standard_answer": "Mitochondrion", "
   answer_format_regex": "Mitochondrion|Mitochondria"}
{"question_id": "FA_BIO_002", "domain": "Factual Accuracy
   ", "question_text_english": "Photosynthesis is the
   process used by plants to convert light energy into
   what type of energy?", "gold_standard_answer": "
   Chemical energy", "answer_format_regex": "Chemical
   energy"}
{"question_id": "FA_BIO_003", "domain": "Factual Accuracy
   ", "question_text_english": "What molecule carries the
   genetic instructions for the development, functioning,
   growth and reproduction of all known organisms and many
    viruses?", "gold_standard_answer": "DNA", "
   answer_format_regex": "DNA|Deoxyribonucleic acid"}
```

```
{"question_id": "FA_BIO_004", "domain": "Factual Accuracy
   ", "question_text_english": "How many chambers are in
   the human heart?", "gold_standard_answer": "Four", "
   answer_format_regex": "Four|4"}
{"question_id": "FA_BIO_005", "domain": "Factual Accuracy
   ", "question_text_english": "What is the scientific
   name for the process of cell division?", "
   gold_standard_answer": "Mitosis", "answer_format_regex
   ": "Mitosis"}
{"question_id": "FA_BIO_006", "domain": "Factual Accuracy
   ", "question_text_english": "Which part of the plant is
    primarily responsible for absorbing water and
   nutrients from the soil?", "gold_standard_answer": "
   Roots", "answer_format_regex": "Roots"}
{"question_id": "FA_BIO_007", "domain": "Factual Accuracy
   ", "question_text_english": "What is the common name
   for the larynx?", "gold_standard_answer": "Voice box",
   "answer_format_regex": "Voice box"}
{"question_id": "FA_BIO_008", "domain": "Factual Accuracy
   ", "question_text_english": "Charles Darwin is famous
   for his theory of evolution by what mechanism?", "
   gold_standard_answer": "Natural selection", "
   answer_format_regex": "Natural selection"}
{"question_id": "FA_BIO_009", "domain": "Factual Accuracy
   ", "question_text_english": "What is the largest organ
   in the human body?", "gold_standard_answer": "Skin", "
   answer_format_regex": "Skin"}
{"question_id": "FA_BIO_010", "domain": "Factual Accuracy
   ", "question_text_english": "What type of animal is a
   mammal?", "gold_standard_answer": "Warm-blooded
   vertebrate", "answer_format_regex": "Warm-blooded
   vertebrate"}
{"question_id": "FA_BIO_011", "domain": "Factual Accuracy
   ", "question_text_english": "What substance gives
   plants their green colour?", "gold_standard_answer": "
   Chlorophyll", "answer_format_regex": "Chlorophyll"}
{"question_id": "FA_BIO_012", "domain": "Factual Accuracy
   ", "question_text_english": "Which blood type is known
   as the universal donor?", "gold_standard_answer": "O
   negative", "answer_format_regex": "O negative|O-"}
{"question_id": "FA_BIO_013", "domain": "Factual Accuracy
   ", "question_text_english": "What are the building
   blocks of proteins?", "gold_standard_answer": "Amino
   acids", "answer_format_regex": "Amino acids"}
{"question_id": "FA_BIO_014", "domain": "Factual Accuracy
   ", "question_text_english": "What is the human body's
   primary source of energy?", "gold_standard_answer": "
   Carbohydrates", "answer_format_regex": "Carbohydrates"}
{"question_id": "FA_BIO_015", "domain": "Factual Accuracy
   ", "question_text_english": "The study of fungi is
   known as what?", "gold_standard_answer": "Mycology", "
   answer_format_regex": "Mycology"}
{"question_id": "FA_BIO_016", "domain": "Factual Accuracy
   ", "question_text_english": "Which bone in the human
   body is the longest?", "gold_standard_answer": "Femur",
    "answer_format_regex": "Femur"}
{"question_id": "FA_BIO_017", "domain": "Factual Accuracy
   ", "question_text_english": "What is the name of the
   process by which bacteria reproduce?", "
```

```
   gold_standard_answer": "Binary fission", "
   answer_format_regex": "Binary fission"}
{"question_id": "FA_BIO_018", "domain": "Factual Accuracy
   ", "question_text_english": "What is the main function
   of the kidneys?", "gold_standard_answer": "Filter blood
    and produce urine", "answer_format_regex": "Filter
   blood|Produce urine"}
{"question_id": "FA_BIO_019", "domain": "Factual Accuracy
   ", "question_text_english": "What type of joint is the
   human shoulder?", "gold_standard_answer": "Ball and
   socket joint", "answer_format_regex": "Ball and socket
   "}
{"question_id": "FA_BIO_020", "domain": "Factual Accuracy
   ", "question_text_english": "Gregor Mendel is known as
   the father of modern genetics for his work on which
   plants?", "gold_standard_answer": "Pea plants", "
   answer_format_regex": "Pea plants|Peas"}
{"question_id": "FA_BIO_021", "domain": "Factual Accuracy
   ", "question_text_english": "What is the largest artery
    in the human body?", "gold_standard_answer": "Aorta",
   "answer_format_regex": "Aorta"}
{"question_id": "FA_BIO_022", "domain": "Factual Accuracy
   ", "question_text_english": "Which part of the brain is
    responsible for balance and coordination?", "
   gold_standard_answer": "Cerebellum", "
   answer_format_regex": "Cerebellum"}
{"question_id": "FA_BIO_023", "domain": "Factual Accuracy
   ", "question_text_english": "What is the process by
   which a caterpillar turns into a butterfly?", "
   gold_standard_answer": "Metamorphosis", "
   answer_format_regex": "Metamorphosis"}
{"question_id": "FA_BIO_024", "domain": "Factual Accuracy
   ", "question_text_english": "Which gas do humans
   primarily exhale when breathing?", "
   gold_standard_answer": "Carbon dioxide", "
   answer_format_regex": "Carbon dioxide|CO2"}
{"question_id": "FA_BIO_025", "domain": "Factual Accuracy
   ", "question_text_english": "What is the name for the
   small sacs in the lungs where gas exchange occurs?", "
   gold_standard_answer": "Alveoli", "answer_format_regex
   ": "Alveoli"}
{"question_id": "FA_BIO_026", "domain": "Factual Accuracy
   ", "question_text_english": "The human skeleton is
   composed of how many bones at birth?", "
   gold_standard_answer": "Around 270", "
   answer_format_regex": "270"}
{"question_id": "FA_BIO_027", "domain": "Factual Accuracy
   ", "question_text_english": "What is the study of birds
    called?", "gold_standard_answer": "Ornithology", "
   answer_format_regex": "Ornithology"}
{"question_id": "FA_BIO_028", "domain": "Factual Accuracy
   ", "question_text_english": "Which of the five senses
   is most closely linked to memory?", "
   gold_standard_answer": "Smell", "answer_format_regex":
   "Smell"}
{"question_id": "FA_BIO_029", "domain": "Factual Accuracy
   ", "question_text_english": "What is the hardest
   substance in the human body?", "gold_standard_answer":
```

```
"Tooth enamel", "answer_format_regex": "Tooth enamel|
Enamel"}
{"question_id": "FA_BIO_030", "domain": "Factual Accuracy
", "question_text_english": "What type of organism is a
 mushroom?", "gold_standard_answer": "Fungus", "
answer_format_regex": "Fungus|Fungi"}
{"question_id": "FA_CHEM_001", "domain": "Factual Accuracy
", "question_text_english": "What is the chemical
symbol for the element gold?", "gold_standard_answer":
"Au", "answer_format_regex": "\\bAu\\b"}
{"question_id": "FA_CHEM_002", "domain": "Factual Accuracy
", "question_text_english": "What is the pH of pure
water at room temperature?", "gold_standard_answer":
"7", "answer_format_regex": "\\b7\\b"}
{"question_id": "FA_CHEM_003", "domain": "Factual Accuracy
", "question_text_english": "What is the most abundant
gas in the Earth's atmosphere?", "gold_standard_answer
": "Nitrogen", "answer_format_regex": "Nitrogen"}
{"question_id": "FA_CHEM_004", "domain": "Factual Accuracy
", "question_text_english": "What is the chemical
formula for table salt?", "gold_standard_answer": "NaCl
", "answer_format_regex": "NaCl"}
{"question_id": "FA_CHEM_005", "domain": "Factual Accuracy
", "question_text_english": "Which element has the
atomic number 1?", "gold_standard_answer": "Hydrogen",
"answer_format_regex": "Hydrogen"}
{"question_id": "FA_CHEM_006", "domain": "Factual Accuracy
", "question_text_english": "The process of a liquid
turning into a gas is called what?", "
gold_standard_answer": "Evaporation", "
answer_format_regex": "Evaporation|Vaporization"}
{"question_id": "FA_CHEM_007", "domain": "Factual Accuracy
", "question_text_english": "What two elements make up
a water molecule?", "gold_standard_answer": "Hydrogen
and Oxygen", "answer_format_regex": "Hydrogen and
Oxygen"}
{"question_id": "FA_CHEM_008", "domain": "Factual Accuracy
", "question_text_english": "What is the common name
for the compound H2O2?", "gold_standard_answer": "
Hydrogen peroxide", "answer_format_regex": "Hydrogen
peroxide"}
{"question_id": "FA_CHEM_009", "domain": "Factual Accuracy
", "question_text_english": "Bronze is an alloy
primarily composed of copper and which other element?",
 "gold_standard_answer": "Tin", "answer_format_regex":
"Tin"}
{"question_id": "FA_CHEM_010", "domain": "Factual Accuracy
", "question_text_english": "Which noble gas is the
lightest?", "gold_standard_answer": "Helium", "
answer_format_regex": "Helium"}
{"question_id": "FA_CHEM_011", "domain": "Factual Accuracy
", "question_text_english": "What is the process called
 when a solid turns directly into a gas?", "
gold_standard_answer": "Sublimation", "
answer_format_regex": "Sublimation"}
{"question_id": "FA_CHEM_012", "domain": "Factual Accuracy
", "question_text_english": "In the periodic table,
what is the symbol for potassium?", "
```

```
   gold_standard_answer": "K", "answer_format_regex": "\\
   bK\\b"}
{"question_id": "FA_CHEM_013", "domain": "Factual Accuracy
   ", "question_text_english": "What type of chemical bond
    involves the sharing of electron pairs?", "
   gold_standard_answer": "Covalent bond", "
   answer_format_regex": "Covalent"}
{"question_id": "FA_CHEM_014", "domain": "Factual Accuracy
   ", "question_text_english": "Acetic acid is the main
   component of what common household liquid, besides
   water?", "gold_standard_answer": "Vinegar", "
   answer_format_regex": "Vinegar"}
{"question_id": "FA_CHEM_015", "domain": "Factual Accuracy
   ", "question_text_english": "Diamonds are a form of
   which element?", "gold_standard_answer": "Carbon", "
   answer_format_regex": "Carbon"}
{"question_id": "FA_CHEM_016", "domain": "Factual Accuracy
   ", "question_text_english": "What is the name of the
   scale used to measure the hardness of a mineral?", "
   gold_standard_answer": "Mohs scale", "
   answer_format_regex": "Mohs"}
{"question_id": "FA_CHEM_017", "domain": "Factual Accuracy
   ", "question_text_english": "What is the term for a
   substance that speeds up a chemical reaction without
   being consumed?", "gold_standard_answer": "Catalyst", "
   answer_format_regex": "Catalyst"}
{"question_id": "FA_CHEM_018", "domain": "Factual Accuracy
   ", "question_text_english": "What is the state of
   matter of mercury at room temperature?", "
   gold_standard_answer": "Liquid", "answer_format_regex":
    "Liquid"}
{"question_id": "FA_CHEM_019", "domain": "Factual Accuracy
   ", "question_text_english": "The nucleus of an atom
   contains which two particles?", "gold_standard_answer":
    "Protons and neutrons", "answer_format_regex": "
   Protons and neutrons"}
{"question_id": "FA_CHEM_020", "domain": "Factual Accuracy
   ", "question_text_english": "The combustion of petrol
   is what type of reaction?", "gold_standard_answer": "
   Exothermic", "answer_format_regex": "Exothermic"}
{"question_id": "FA_CHEM_021", "domain": "Factual Accuracy
   ", "question_text_english": "What element is
   represented by the chemical symbol Fe?", "
   gold_standard_answer": "Iron", "answer_format_regex": "
   Iron"}
{"question_id": "FA_CHEM_022", "domain": "Factual Accuracy
   ", "question_text_english": "What is the primary
   element that makes up organic molecules?", "
   gold_standard_answer": "Carbon", "answer_format_regex":
    "Carbon"}
{"question_id": "FA_CHEM_023", "domain": "Factual Accuracy
   ", "question_text_english": "Ozone is an allotrope of
   which element?", "gold_standard_answer": "Oxygen", "
   answer_format_regex": "Oxygen"}
{"question_id": "FA_CHEM_024", "domain": "Factual Accuracy
   ", "question_text_english": "What is the chemical
   formula for methane?", "gold_standard_answer": "CH4", "
   answer_format_regex": "CH4"}
```

```
{"question_id": "FA_CHEM_025", "domain": "Factual Accuracy
   ", "question_text_english": "What term describes a
   solution with a pH greater than 7?", "
   gold_standard_answer": "Alkaline", "answer_format_regex
   ": "Alkaline|Basic"}
{"question_id": "FA_CHEM_026", "domain": "Factual Accuracy
   ", "question_text_english": "Which is the most
   electronegative element?", "gold_standard_answer": "
   Fluorine", "answer_format_regex": "Fluorine"}
{"question_id": "FA_CHEM_027", "domain": "Factual Accuracy
   ", "question_text_english": "What is the chemical
   symbol for silver?", "gold_standard_answer": "Ag", "
   answer_format_regex": "\\bAg\\b"}
{"question_id": "FA_CHEM_028", "domain": "Factual Accuracy
   ", "question_text_english": "The process of splitting a
    heavy atomic nucleus is known as what?", "
   gold_standard_answer": "Nuclear fission", "
   answer_format_regex": "Fission"}
{"question_id": "FA_CHEM_029", "domain": "Factual Accuracy
   ", "question_text_english": "What is the common name
   for solid carbon dioxide?", "gold_standard_answer": "
   Dry ice", "answer_format_regex": "Dry ice"}
{"question_id": "FA_CHEM_030", "domain": "Factual Accuracy
   ", "question_text_english": "Which gas is produced
   during photosynthesis?", "gold_standard_answer": "
   Oxygen", "answer_format_regex": "Oxygen"}
{"question_id": "FA_PHYS_001", "domain": "Factual Accuracy
   ", "question_text_english": "What is the unit of
   electrical resistance?", "gold_standard_answer": "Ohm",
    "answer_format_regex": "Ohm"}
{"question_id": "FA_PHYS_002", "domain": "Factual Accuracy
   ", "question_text_english": "Sir Isaac Newton is famous
    for his three laws of what?", "gold_standard_answer":
   "Motion", "answer_format_regex": "Motion"}
{"question_id": "FA_PHYS_003", "domain": "Factual Accuracy
   ", "question_text_english": "What is the speed of light
    in a vacuum, commonly denoted by 'c'?", "
   gold_standard_answer": "299,792,458 meters per second",
    "answer_format_regex": "299,792,458|3\\.00 x 10\\^8"}
{"question_id": "FA_PHYS_004", "domain": "Factual Accuracy
   ", "question_text_english": "What force is responsible
   for keeping the planets in orbit around the Sun?", "
   gold_standard_answer": "Gravity", "answer_format_regex
   ": "Gravity"}
{"question_id": "FA_PHYS_005", "domain": "Factual Accuracy
   ", "question_text_english": "What is the standard unit
   of mass in the International System of Units (SI)?", "
   gold_standard_answer": "Kilogram", "answer_format_regex
   ": "Kilogram|kg"}
{"question_id": "FA_PHYS_006", "domain": "Factual Accuracy
   ", "question_text_english": "The tendency of an object
   to resist a change in its state of motion is called
   what?", "gold_standard_answer": "Inertia", "
   answer_format_regex": "Inertia"}
{"question_id": "FA_PHYS_007", "domain": "Factual Accuracy
   ", "question_text_english": "What is the unit of
   frequency, equivalent to one cycle per second?", "
   gold_standard_answer": "Hertz", "answer_format_regex":
   "Hertz|Hz"}
```

{"question_id": "FA_PHYS_008", "domain": "Factual Accuracy", "question_text_english": "What type of energy is stored in an object due to its position in a gravitational field?", "gold_standard_answer": "Potential energy", "answer_format_regex": "Potential energy"}
{"question_id": "FA_PHYS_009", "domain": "Factual Accuracy", "question_text_english": "Albert Einstein's famous equation, E=mc\u00b2, relates energy to mass and what other quantity?", "gold_standard_answer": "The speed of light", "answer_format_regex": "The speed of light|Speed of light"}
{"question_id": "FA_PHYS_010", "domain": "Factual Accuracy", "question_text_english": "What is the boiling point of water at standard atmospheric pressure in degrees Celsius?", "gold_standard_answer": "100", "answer_format_regex": "\\b100\\b"}
{"question_id": "FA_PHYS_011", "domain": "Factual Accuracy", "question_text_english": "Which colour has the longest wavelength in the visible spectrum?", "gold_standard_answer": "Red", "answer_format_regex": "Red"}
{"question_id": "FA_PHYS_012", "domain": "Factual Accuracy", "question_text_english": "What is the term for the flow of electric charge?", "gold_standard_answer": "Electric current", "answer_format_regex": "Electric current|Current"}
{"question_id": "FA_PHYS_013", "domain": "Factual Accuracy", "question_text_english": "Sound waves travel fastest through which state of matter?", "gold_standard_answer": "Solid", "answer_format_regex": "Solid|Solids"}
{"question_id": "FA_PHYS_014", "domain": "Factual Accuracy", "question_text_english": "Who is credited with the discovery of the electron?", "gold_standard_answer": "J.J. Thomson", "answer_format_regex": "J\\.J\\. Thomson|Thomson"}
{"question_id": "FA_PHYS_015", "domain": "Factual Accuracy", "question_text_english": "What is the common name for a device that converts mechanical energy into electrical energy?", "gold_standard_answer": "Generator", "answer_format_regex": "Generator"}
{"question_id": "FA_PHYS_016", "domain": "Factual Accuracy", "question_text_english": "What law states that for every action, there is an equal and opposite reaction?", "gold_standard_answer": "Newton's Third Law of Motion", "answer_format_regex": "Newton's Third Law"}
{"question_id": "FA_PHYS_017", "domain": "Factual Accuracy", "question_text_english": "The measure of a material's ability to transmit light is called what?", "gold_standard_answer": "Transparency", "answer_format_regex": "Transparency"}
{"question_id": "FA_PHYS_018", "domain": "Factual Accuracy", "question_text_english": "What is the standard unit of power, equivalent to one joule per second?", "gold_standard_answer": "Watt", "answer_format_regex": "Watt|W"}
{"question_id": "FA_PHYS_019", "domain": "Factual Accuracy", "question_text_english": "In a standard electrical outlet in the UK, what is the nominal voltage?", "

```
   gold_standard_answer": "230 volts", "
   answer_format_regex": "230|230V"}
{"question_id": "FA_PHYS_020", "domain": "Factual Accuracy
   ", "question_text_english": "What phenomenon causes a
   rainbow?", "gold_standard_answer": "Refraction and
   dispersion of light", "answer_format_regex": "
   Refraction|Dispersion"}
{"question_id": "FA_PHYS_021", "domain": "Factual Accuracy
   ", "question_text_english": "What is the unit used to
   measure electric current?", "gold_standard_answer": "
   Ampere", "answer_format_regex": "Ampere|Amp|A"}
{"question_id": "FA_PHYS_022", "domain": "Factual Accuracy
   ", "question_text_english": "What type of simple
   machine is a doorknob?", "gold_standard_answer": "Wheel
    and axle", "answer_format_regex": "Wheel and axle"}
{"question_id": "FA_PHYS_023", "domain": "Factual Accuracy
   ", "question_text_english": "What is the freezing point
    of water in degrees Celsius?", "gold_standard_answer":
    "0", "answer_format_regex": "\\b0\\b"}
{"question_id": "FA_PHYS_024", "domain": "Factual Accuracy
   ", "question_text_english": "Which of Newton's laws is
   also known as the law of inertia?", "
   gold_standard_answer": "First Law", "
   answer_format_regex": "First|1st"}
{"question_id": "FA_PHYS_025", "domain": "Factual Accuracy
   ", "question_text_english": "What is the term for a
   material that does not allow electricity to pass
   through it?", "gold_standard_answer": "Insulator", "
   answer_format_regex": "Insulator"}
{"question_id": "FA_PHYS_026", "domain": "Factual Accuracy
   ", "question_text_english": "What form of heat transfer
    is responsible for the sun warming the earth?", "
   gold_standard_answer": "Radiation", "
   answer_format_regex": "Radiation"}
{"question_id": "FA_PHYS_027", "domain": "Factual Accuracy
   ", "question_text_english": "What is the name of the
   particle that carries the strong nuclear force?", "
   gold_standard_answer": "Gluon", "answer_format_regex":
   "Gluon"}
{"question_id": "FA_PHYS_028", "domain": "Factual Accuracy
   ", "question_text_english": "The decibel (dB) is a unit
    used to measure the intensity of what?", "
   gold_standard_answer": "Sound", "answer_format_regex":
   "Sound"}
{"question_id": "FA_PHYS_029", "domain": "Factual Accuracy
   ", "question_text_english": "What is the term for the
   bending of light as it passes from one medium to
   another?", "gold_standard_answer": "Refraction", "
   answer_format_regex": "Refraction"}
{"question_id": "FA_PHYS_030", "domain": "Factual Accuracy
   ", "question_text_english": "Which planet in our solar
   system is known for its prominent rings?", "
   gold_standard_answer": "Saturn", "answer_format_regex":
    "Saturn"}
```

B Procedural Reasoning Questions

The following table is a set of questions obtained from the GSM8K question set. Each question has then been evaluated and where necessary corrected to the EN(UK) equivalent.

Proposed Question Corpus: Procedural Reasoning (200 Questions)


```
{"question_id": "LR_MATH_001", "domain": "Procedural
   Reasoning", "question_text_english": "A baker has 12
   dozen eggs. He uses 3 dozen for a large cake. How many
   individual eggs does he have left?", "
   gold_standard_answer": "108", "answer_format_regex":
   "\\b108\\b", "gold_standard_reasoning": "Step 1:
   Calculate total eggs. 12 dozen * 12 eggs/dozen = 144
   eggs. Step 2: Calculate eggs used. 3 dozen * 12 eggs/
   dozen = 36 eggs. Step 3: Subtract used from total. 144
   - 36 = 108 eggs left."}
{"question_id": "LR_MATH_002", "domain": "Procedural
   Reasoning", "question_text_english": "A library has 25
   shelves, and each shelf can hold 40 books. If 3 shelves
    are empty, how many books are in the library?", "
   gold_standard_answer": "880", "answer_format_regex":
   "\\b880\\b", "gold_standard_reasoning": "Step 1:
   Calculate the number of full shelves. 25 shelves - 3
   empty shelves = 22 full shelves. Step 2: Calculate the
   total number of books. 22 shelves * 40 books/shelf =
   880 books."}
{"question_id": "LR_MATH_003", "domain": "Procedural
   Reasoning", "question_text_english": "John buys a toy
   for \u00a38.50 and a chocolate bar for \u00a31.25. He
   pays with a \u00a320 note. How much change does he
   receive?", "gold_standard_answer": "9.25", "
   answer_format_regex": "9\\.25|\u00a39\\.25", "
   gold_standard_reasoning": "Step 1: Calculate the total
   cost. \u00a38.50 + \u00a31.25 = \u00a39.75. Step 2:
   Calculate the change. \u00a320.00 - \u00a39.75 = \
   u00a39.25."}
{"question_id": "LR_MATH_004", "domain": "Procedural
   Reasoning", "question_text_english": "A train journey
   from London to Manchester is 200 miles. If the train
   travels at an average speed of 80 mph, how many hours
   will the journey take?", "gold_standard_answer": "2.5",
    "answer_format_regex": "2\\.5", "
   gold_standard_reasoning": "Step 1: Use the formula Time
    = Distance / Speed. Step 2: Calculate the time. 200
   miles / 80 mph = 2.5 hours."}
{"question_id": "LR_MATH_005", "domain": "Procedural
   Reasoning", "question_text_english": "Sarah has 3 bags
   of marbles. Each bag contains 15 red marbles and 10
   blue marbles. How many marbles does she have in total
   ?", "gold_standard_answer": "75", "answer_format_regex
   ": "\\b75\\b", "gold_standard_reasoning": "Step 1:
   Calculate the number of marbles per bag. 15 red + 10
   blue = 25 marbles. Step 2: Calculate the total number
   of marbles. 3 bags * 25 marbles/bag = 75 marbles."}
{"question_id": "LR_MATH_006", "domain": "Procedural
   Reasoning", "question_text_english": "A farmer has 50
   chickens. Each chicken lays 4 eggs a week. How many
   eggs will the farmer collect in 3 weeks?", "
   gold_standard_answer": "600", "answer_format_regex":
```

```
   "\\b600\\b", "gold_standard_reasoning": "Step 1:
   Calculate the number of eggs per week. 50 chickens * 4
   eggs/chicken = 200 eggs per week. Step 2: Calculate the
    total eggs for 3 weeks. 200 eggs/week * 3 weeks = 600
   eggs."}
{"question_id": "LR_MATH_007", "domain": "Procedural
   Reasoning", "question_text_english": "A recipe calls
   for 250 grams of flour to make 12 biscuits. How many
   grams of flour are needed to make 30 biscuits?", "
   gold_standard_answer": "625", "answer_format_regex":
   "\\b625\\b", "gold_standard_reasoning": "Step 1:
   Calculate the flour needed per biscuit. 250 grams / 12
   biscuits = 20.833 g/biscuit. Step 2: Calculate the
   total flour needed for 30 biscuits. 30 biscuits *
   20.833 g/biscuit = 625 grams. Alternate method: (30/12)
    * 250 = 2.5 * 250 = 625."}
{"question_id": "LR_MATH_008", "domain": "Procedural
   Reasoning", "question_text_english": "A car's petrol
   tank holds 60 litres. If it gets 10 miles per litre,
   how many miles can it travel on a full tank?", "
   gold_standard_answer": "600", "answer_format_regex":
   "\\b600\\b", "gold_standard_reasoning": "Step 1: Use
   the formula Distance = Fuel Capacity * Efficiency. Step
    2: Calculate the total distance. 60 litres * 10 miles/
   litre = 600 miles."}
{"question_id": "LR_MATH_009", "domain": "Procedural
   Reasoning", "question_text_english": "A school has 450
   pupils. 2/5 of the pupils are boys. How many girls are
   there in the school?", "gold_standard_answer": "270", "
   answer_format_regex": "\\b270\\b", "
   gold_standard_reasoning": "Step 1: Calculate the number
    of boys. 450 pupils * (2/5) = 180 boys. Step 2:
   Calculate the number of girls by subtracting the boys
   from the total. 450 total pupils - 180 boys = 270 girls
   ."}
{"question_id": "LR_MATH_010", "domain": "Procedural
   Reasoning", "question_text_english": "A film starts at
   18:45 and is 1 hour and 50 minutes long. What time does
    it finish?", "gold_standard_answer": "20:35", "
   answer_format_regex": "20:35|8:35 PM", "
   gold_standard_reasoning": "Step 1: Add the hours. 18:45
    + 1 hour = 19:45. Step 2: Add the minutes. 19:45 + 50
   minutes. 45 + 50 = 95 minutes. Step 3: Convert extra
   minutes to hours. 95 minutes = 1 hour and 35 minutes.
   Step 4: Add this to the time. 19:00 + 1 hour and 35
   minutes = 20:35."}
{"question_id": "LR_MATH_011", "domain": "Procedural
   Reasoning", "question_text_english": "A shopkeeper buys
    a box of 48 apples for \u00a312. He sells them for 50
   pence each. How much profit does he make?", "
   gold_standard_answer": "12", "answer_format_regex": "\\
   b12\\b|\u00a312", "gold_standard_reasoning": "Step 1:
   Calculate total revenue. 48 apples * 50 pence/apple =
   2400 pence. Step 2: Convert revenue to pounds. 2400
   pence = \u00a324. Step 3: Calculate the profit. \
   u00a324 (revenue) - \u00a312 (cost) = \u00a312 profit
   ."}
{"question_id": "LR_MATH_012", "domain": "Procedural
   Reasoning", "question_text_english": "A rectangular
```

garden is 15 metres long and 8 metres wide. What is its perimeter?", "gold_standard_answer": "46", "answer_format_regex": "\\b46\\b", "gold_standard_reasoning": "Step 1: Use the formula for the perimeter of a rectangle: P = 2 * (length + width). Step 2: Substitute the values. P = 2 * (15 + 8). Step 3: Calculate the result. P = 2 * 23 = 46 metres."}
{"question_id": "LR_MATH_013", "domain": "Procedural Reasoning", "question_text_english": "If a pack of 6 pens costs \u00a34.20, what is the cost of one pen in pence?", "gold_standard_answer": "70", "answer_format_regex": "\\b70\\b", "gold_standard_reasoning": "Step 1: Convert the total cost to pence. \u00a34.20 = 420 pence. Step 2: Calculate the cost per pen. 420 pence / 6 pens = 70 pence."}
{"question_id": "LR_MATH_014", "domain": "Procedural Reasoning", "question_text_english": "A bus travels 60 km in one hour. How many kilometres will it travel in 2.5 hours?", "gold_standard_answer": "150", "answer_format_regex": "\\b150\\b", "gold_standard_reasoning": "Step 1: Use the formula Distance = Speed * Time. Step 2: The speed is 60 km/h. The time is 2.5 hours. Step 3: Calculate the distance. 60 km/h * 2.5 h = 150 kilometres."}
{"question_id": "LR_MATH_015", "domain": "Procedural Reasoning", "question_text_english": "There are 30 days in September. If it rains on 12 of those days, what percentage of days were not rainy?", "gold_standard_answer": "60", "answer_format_regex": "60|60%", "gold_standard_reasoning": "Step 1: Calculate the number of non-rainy days. 30 days - 12 rainy days = 18 non-rainy days. Step 2: Calculate the percentage. (18 non-rainy days / 30 total days) * 100 = 60%."}
{"question_id": "LR_MATH_016", "domain": "Procedural Reasoning", "question_text_english": "A class has 28 pupils. 3/4 of the pupils passed their exam. How many pupils failed?", "gold_standard_answer": "7", "answer_format_regex": "\\b7\\b", "gold_standard_reasoning": "Step 1: The fraction of pupils who failed is 1 - 3/4 = 1/4. Step 2: Calculate the number of pupils who failed. 28 pupils * (1/4) = 7 pupils."}
{"question_id": "LR_MATH_017", "domain": "Procedural Reasoning", "question_text_english": "A book has 320 pages. Tom has read 1/4 of the book on Monday and 1/2 of the remaining pages on Tuesday. How many pages are left to read?", "gold_standard_answer": "120", "answer_format_regex": "\\b120\\b", "gold_standard_reasoning": "Step 1: Pages read on Monday. 320 * (1/4) = 80 pages. Step 2: Remaining pages after Monday. 320 - 80 = 240 pages. Step 3: Pages read on Tuesday. 240 * (1/2) = 120 pages. Step 4: Pages left to read. 240 - 120 = 120 pages."}
{"question_id": "LR_MATH_018", "domain": "Procedural Reasoning", "question_text_english": "A box contains 24 cans of fizzy drink. If a single can weighs 350 grams, what is the total weight of the cans in the box, in kilograms?", "gold_standard_answer": "8.4", "

```
   answer_format_regex": "8\\.4", "gold_standard_reasoning
   ": "Step 1: Calculate the total weight in grams. 24
   cans * 350 grams/can = 8400 grams. Step 2: Convert
   grams to kilograms. 8400 grams / 1000 g/kg = 8.4
   kilograms."}
{"question_id": "LR_MATH_019", "domain": "Procedural
   Reasoning", "question_text_english": "A cyclist is
   travelling at 15 miles per hour. How many minutes will
   it take them to travel 5 miles?", "gold_standard_answer
   ": "20", "answer_format_regex": "\\b20\\b", "
   gold_standard_reasoning": "Step 1: Calculate the time
   in hours. Time = Distance / Speed = 5 miles / 15 mph =
   1/3 hours. Step 2: Convert hours to minutes. (1/3)
   hours * 60 minutes/hour = 20 minutes."}
{"question_id": "LR_MATH_020", "domain": "Procedural
   Reasoning", "question_text_english": "A supermarket has
    8 aisles. Each aisle has 2 sides, and each side has 10
    shelves. How many shelves are there in total?", "
   gold_standard_answer": "160", "answer_format_regex":
   "\\b160\\b", "gold_standard_reasoning": "Step 1:
   Calculate shelves per aisle. 2 sides * 10 shelves/side
   = 20 shelves per aisle. Step 2: Calculate total shelves
   . 8 aisles * 20 shelves/aisle = 160 shelves."}
{"question_id": "LR_MATH_021", "domain": "Procedural
   Reasoning", "question_text_english": "A phone's battery
    is at 20%. If it charges 10% every 5 minutes, how many
    minutes will it take to be fully charged?", "
   gold_standard_answer": "40", "answer_format_regex": "\\
   b40\\b", "gold_standard_reasoning": "Step 1: Calculate
   the percentage needed to charge. 100% - 20% = 80%. Step
    2: Calculate the number of 10% increments needed. 80%
   / 10% = 8 increments. Step 3: Calculate the total time.
    8 increments * 5 minutes/increment = 40 minutes."}
{"question_id": "LR_MATH_022", "domain": "Procedural
   Reasoning", "question_text_english": "There are 5 red
   balls, 3 blue balls, and 2 green balls in a bag. What
   is the probability of picking a blue ball?", "
   gold_standard_answer": "0.3", "answer_format_regex":
   "0\\.3|3/10|30%", "gold_standard_reasoning": "Step 1:
   Calculate the total number of balls. 5 + 3 + 2 = 10
   balls. Step 2: Identify the number of favourable
   outcomes (blue balls), which is 3. Step 3: Calculate
   the probability. Probability = (Favourable Outcomes) /
   (Total Outcomes) = 3 / 10 = 0.3."}
{"question_id": "LR_MATH_023", "domain": "Procedural
   Reasoning", "question_text_english": "A temperature of
   25 degrees Celsius is how many degrees Fahrenheit? (
   Formula: F = C * 9/5 + 32)", "gold_standard_answer":
   "77", "answer_format_regex": "\\b77\\b", "
   gold_standard_reasoning": "Step 1: Substitute C=25 into
    the formula. F = 25 * (9/5) + 32. Step 2: Perform the
   multiplication. 25 * 1.8 = 45. Step 3: Perform the
   addition. 45 + 32 = 77."}
{"question_id": "LR_MATH_024", "domain": "Procedural
   Reasoning", "question_text_english": "A piece of wood
   is 2.4 metres long. It is cut into 3 equal pieces. How
   long is each piece in centimetres?", "
   gold_standard_answer": "80", "answer_format_regex": "\\
   b80\\b", "gold_standard_reasoning": "Step 1: Convert
```

the total length to centimetres. 2.4 metres * 100 cm/m = 240 cm. Step 2: Divide by the number of pieces. 240 cm / 3 = 80 cm."}
{"question_id": "LR_MATH_025", "domain": "Procedural Reasoning", "question_text_english": "A shop offers a 20% discount on a TV that costs \u00a3500. What is the final price?", "gold_standard_answer": "400", "answer_format_regex": "\\b400\\b|\u00a3400", "gold_standard_reasoning": "Step 1: Calculate the discount amount. \u00a3500 * 20% = \u00a3100. Step 2: Subtract the discount from the original price. \u00a3500 - \u00a3100 = \u00a3400."}
{"question_id": "LR_MATH_026", "domain": "Procedural Reasoning", "question_text_english": "In a class of 30 pupils, the ratio of boys to girls is 2:3. How many boys are there?", "gold_standard_answer": "12", "answer_format_regex": "\\b12\\b", "gold_standard_reasoning": "Step 1: Find the total number of parts in the ratio. 2 + 3 = 5 parts. Step 2: Calculate the value of one part. 30 pupils / 5 parts = 6 pupils per part. Step 3: Calculate the number of boys. 2 parts * 6 pupils/part = 12 boys."}
{"question_id": "LR_MATH_027", "domain": "Procedural Reasoning", "question_text_english": "A tap fills a 100-litre bath in 10 minutes. What is the flow rate in litres per minute?", "gold_standard_answer": "10", "answer_format_regex": "\\b10\\b", "gold_standard_reasoning": "Step 1: Use the formula Flow Rate = Volume / Time. Step 2: Substitute the values. Flow Rate = 100 litres / 10 minutes. Step 3: Calculate the result. 10 litres per minute."}
{"question_id": "LR_MATH_028", "domain": "Procedural Reasoning", "question_text_english": "A writer types at 60 words per minute. How many words can they type in 1.5 hours?", "gold_standard_answer": "5400", "answer_format_regex": "\\b5400\\b", "gold_standard_reasoning": "Step 1: Convert the time to minutes. 1.5 hours * 60 minutes/hour = 90 minutes. Step 2: Calculate the total words typed. 90 minutes * 60 words/minute = 5400 words."}
{"question_id": "LR_MATH_029", "domain": "Procedural Reasoning", "question_text_english": "The sum of three consecutive even numbers is 42. What is the smallest of these numbers?", "gold_standard_answer": "12", "answer_format_regex": "\\b12\\b", "gold_standard_reasoning": "Step 1: Let the numbers be x, x+2, and x+4. Step 2: Set up the equation. x + (x+2) + (x+4) = 42. Step 3: Solve for x. 3x + 6 = 42. 3x = 36. x = 12. The smallest number is 12."}
{"question_id": "LR_MATH_030", "domain": "Procedural Reasoning", "question_text_english": "A packet of crisps costs 80 pence. A multipack of 6 packets costs \u00a33.60. How much do you save per packet by buying the multipack, in pence?", "gold_standard_answer": "20", "answer_format_regex": "\\b20\\b", "gold_standard_reasoning": "Step 1: Calculate the cost per packet in the multipack. \u00a33.60 = 360 pence. 360 pence / 6 packets = 60 pence per packet. Step 2:

```
   Calculate the saving per packet. 80 pence - 60 pence =
   20 pence."}
{"question_id": "LR_MATH_031", "domain": "Procedural
   Reasoning", "question_text_english": "A car travels at
   50 mph for 2 hours and then at 70 mph for 1 hour. What
   is the total distance travelled?", "
   gold_standard_answer": "170", "answer_format_regex":
   "\\b170\\b", "gold_standard_reasoning": "Step 1:
   Calculate the distance for the first part of the
   journey. 50 mph * 2 hours = 100 miles. Step 2:
   Calculate the distance for the second part. 70 mph * 1
   hour = 70 miles. Step 3: Add the distances together.
   100 miles + 70 miles = 170 miles."}
{"question_id": "LR_MATH_032", "domain": "Procedural
   Reasoning", "question_text_english": "A square has an
   area of 64 square centimetres. What is the length of
   one of its sides?", "gold_standard_answer": "8", "
   answer_format_regex": "\\b8\\b", "
   gold_standard_reasoning": "Step 1: The area of a square
    is the side length squared (A = s^2). Step 2: To find
   the side length, take the square root of the area. s =
   sqrt(64). Step 3: The square root of 64 is 8. So the
   side length is 8 cm."}
{"question_id": "LR_MATH_033", "domain": "Procedural
   Reasoning", "question_text_english": "If a company's
   profit was \u00a31.2 million last year and increased by
    15% this year, what is this year's profit?", "
   gold_standard_answer": "1380000", "answer_format_regex
   ": "1380000|1\\.38 million|\u00a31\\.38m", "
   gold_standard_reasoning": "Step 1: Calculate the profit
    increase. \u00a31,200,000 * 15% = \u00a3180,000. Step
   2: Add the increase to last year's profit. \u00a31
   ,200,000 + \u00a3180,000 = \u00a31,380,000."}
{"question_id": "LR_MATH_034", "domain": "Procedural
   Reasoning", "question_text_english": "A baker makes 180
    loaves of bread. He sells 5/6 of them. How many loaves
    are left?", "gold_standard_answer": "30", "
   answer_format_regex": "\\b30\\b", "
   gold_standard_reasoning": "Step 1: Calculate the number
    of loaves sold. 180 loaves * (5/6) = 150 loaves sold.
   Step 2: Calculate the number of loaves left. 180 total
   loaves - 150 sold loaves = 30 loaves left."}
{"question_id": "LR_MATH_035", "domain": "Procedural
   Reasoning", "question_text_english": "A work team of 5
   people can complete a job in 8 days. How many days
   would it take a team of 4 people to complete the same
   job, assuming they work at the same rate?", "
   gold_standard_answer": "10", "answer_format_regex": "\\
   b10\\b", "gold_standard_reasoning": "Step 1: Calculate
   the total person-days required for the job. 5 people *
   8 days = 40 person-days. Step 2: Calculate the time for
    the new team. 40 person-days / 4 people = 10 days."}
{"question_id": "LR_MATH_036", "domain": "Procedural
   Reasoning", "question_text_english": "A concert ticket
   costs \u00a360 plus a 10% booking fee. What is the
   total cost of the ticket?", "gold_standard_answer":
   "66", "answer_format_regex": "\\b66\\b|\u00a366", "
   gold_standard_reasoning": "Step 1: Calculate the
   booking fee. \u00a360 * 10% = \u00a36. Step 2: Add the
```

```
   fee to the ticket price. \u00a360 + \u00a36 = \u00a366
   ."}
{"question_id": "LR_MATH_037", "domain": "Procedural
   Reasoning", "question_text_english": "A runner
   completes a 10-kilometre race in 50 minutes. What was
   their average speed in kilometres per hour?", "
   gold_standard_answer": "12", "answer_format_regex": "\\
   b12\\b", "gold_standard_reasoning": "Step 1: Convert
   the time to hours. 50 minutes / 60 minutes/hour = 5/6
   hours. Step 2: Use the formula Speed = Distance / Time.
    Speed = 10 km / (5/6) hours. Step 3: Calculate the
   result. Speed = 10 * (6/5) = 12 km/h."}
{"question_id": "LR_MATH_038", "domain": "Procedural
   Reasoning", "question_text_english": "An angle is 30
   degrees. What is the size of its complementary angle?",
    "gold_standard_answer": "60", "answer_format_regex":
   "\\b60\\b", "gold_standard_reasoning": "Step 1:
   Complementary angles add up to 90 degrees. Step 2:
   Subtract the given angle from 90. 90 degrees - 30
   degrees = 60 degrees."}
{"question_id": "LR_MATH_039", "domain": "Procedural
   Reasoning", "question_text_english": "A farmer plants 5
    rows of trees. Each row has 12 apple trees and 8 pear
   trees. How many trees did he plant in total?", "
   gold_standard_answer": "100", "answer_format_regex":
   "\\b100\\b", "gold_standard_reasoning": "Step 1:
   Calculate the number of trees per row. 12 apple trees +
    8 pear trees = 20 trees per row. Step 2: Calculate the
    total number of trees. 5 rows * 20 trees/row = 100
   trees."}
{"question_id": "LR_MATH_040", "domain": "Procedural
   Reasoning", "question_text_english": "A box of cereal
   costs \u00a33. If a family eats 2 boxes a week, how
   much will they spend on cereal in a year (52 weeks)?",
   "gold_standard_answer": "312", "answer_format_regex":
   "\\b312\\b|\u00a3312", "gold_standard_reasoning": "Step
    1: Calculate the weekly cost. \u00a33/box * 2 boxes =
   \u00a36 per week. Step 2: Calculate the annual cost. \
   u00a36/week * 52 weeks = \u00a3312."}
{"question_id": "LR_MATH_041", "domain": "Procedural
   Reasoning", "question_text_english": "What is 25% of
   300?", "gold_standard_answer": "75", "
   answer_format_regex": "\\b75\\b", "
   gold_standard_reasoning": "Step 1: Convert the
   percentage to a decimal. 25% = 0.25. Step 2: Multiply
   the number by the decimal. 300 * 0.25 = 75."}
{"question_id": "LR_MATH_042", "domain": "Procedural
   Reasoning", "question_text_english": "A map has a scale
    of 1:50,000. If two towns are 4 cm apart on the map,
   what is the real distance between them in kilometres?",
    "gold_standard_answer": "2", "answer_format_regex":
   "\\b2\\b", "gold_standard_reasoning": "Step 1:
   Calculate the real distance in cm. 4 cm * 50,000 =
   200,000 cm. Step 2: Convert cm to metres. 200,000 cm /
   100 cm/m = 2,000 m. Step 3: Convert metres to
   kilometres. 2,000 m / 1000 m/km = 2 km."}
{"question_id": "LR_MATH_043", "domain": "Procedural
   Reasoning", "question_text_english": "A train has 8
   carriages, and each carriage has 64 seats. If 450
```

```
people are on the train, how many empty seats are there
?", "gold_standard_answer": "62", "answer_format_regex
": "\\b62\\b", "gold_standard_reasoning": "Step 1:
Calculate the total number of seats. 8 carriages * 64
seats/carriage = 512 seats. Step 2: Calculate the
number of empty seats. 512 total seats - 450 people =
62 empty seats."}
{"question_id": "LR_MATH_044", "domain": "Procedural
Reasoning", "question_text_english": "A cube has a side
 length of 5 cm. What is its volume?", "
gold_standard_answer": "125", "answer_format_regex":
"\\b125\\b", "gold_standard_reasoning": "Step 1: The
volume of a cube is side length cubed (V = s^3). Step
2: Calculate the volume. 5 * 5 * 5 = 125 cubic cm."}
{"question_id": "LR_MATH_045", "domain": "Procedural
Reasoning", "question_text_english": "A pizza is cut
into 8 equal slices. If Mark eats 3 slices, what
fraction of the pizza is left?", "gold_standard_answer
": "5/8", "answer_format_regex": "5/8", "
gold_standard_reasoning": "Step 1: The whole pizza is
8/8. Step 2: Subtract the eaten portion from the whole.
 8/8 - 3/8 = 5/8. So, 5/8 of the pizza is left."}
{"question_id": "LR_MATH_046", "domain": "Procedural
Reasoning", "question_text_english": "A phone call
costs 15 pence per minute. How much does a 12-minute
call cost in pounds?", "gold_standard_answer": "1.80",
"answer_format_regex": "1\\\\.80|1\\\\.8|\u00a31
\\\\.80", "gold_standard_reasoning": "Step 1: Calculate
 the total cost in pence. 15 pence/minute * 12 minutes
= 180 pence. Step 2: Convert pence to pounds. 180 pence
 / 100 pence/\u00a3 = \u00a31.80."}
{"question_id": "LR_MATH_047", "domain": "Procedural
Reasoning", "question_text_english": "A swimming pool
is 25 metres long, 10 metres wide, and 2 metres deep.
What is the volume of the pool in cubic metres?", "
gold_standard_answer": "500", "answer_format_regex":
"\\b500\\b", "gold_standard_reasoning": "Step 1: Use
the formula for the volume of a rectangular prism: V =
length * width * depth. Step 2: Substitute the values.
V = 25 * 10 * 2. Step 3: Calculate the result. V = 500
cubic metres."}
{"question_id": "LR_MATH_048", "domain": "Procedural
Reasoning", "question_text_english": "If you save \
u00a350 a week, how many weeks will it take to save \
u00a31,200?", "gold_standard_answer": "24", "
answer_format_regex": "\\b24\\b", "
gold_standard_reasoning": "Step 1: Use the formula Time
 = Total Amount / Rate. Step 2: Substitute the values.
Time = \u00a31200 / \u00a350 per week. Step 3:
Calculate the result. 24 weeks."}
{"question_id": "LR_MATH_049", "domain": "Procedural
Reasoning", "question_text_english": "A circle has a
radius of 7 cm. What is its area? (Use pi \u2248 3.14)
", "gold_standard_answer": "153.86", "
answer_format_regex": "153\\\\.86", "
gold_standard_reasoning": "Step 1: Use the formula for
the area of a circle: A = pi * r^2. Step 2: Substitute
the values. A = 3.14 * (7^2). Step 3: Calculate the
result. A = 3.14 * 49 = 153.86 square cm."}
```

```
{"question_id": "LR_MATH_050", "domain": "Procedural
   Reasoning", "question_text_english": "A test has 50
   questions. For every correct answer, you get 2 points.
   For every incorrect answer, you lose 1 point. If a
   student answers all questions and gets 35 correct, what
    is their final score?", "gold_standard_answer": "55",
   "answer_format_regex": "\\b55\\b", "
   gold_standard_reasoning": "Step 1: Calculate points
   from correct answers. 35 correct * 2 points/correct =
   70 points. Step 2: Calculate the number of incorrect
   answers. 50 total - 35 correct = 15 incorrect. Step 3:
   Calculate points lost from incorrect answers. 15
   incorrect * -1 point/incorrect = -15 points. Step 4:
   Calculate the final score. 70 - 15 = 55 points."}
{"question_id": "LR_MATH_051", "domain": "Procedural
   Reasoning", "question_text_english": "A car's price is
   reduced from \u00a325,000 to \u00a322,000. What is the
   percentage discount?", "gold_standard_answer": "12", "
   answer_format_regex": "12|12%", "
   gold_standard_reasoning": "Step 1: Calculate the
   discount amount. \u00a325,000 - \u00a322,000 = \u00a33
   ,000. Step 2: Calculate the percentage discount. (\
   u00a33,000 / \u00a325,000) * 100 = 12%."}
{"question_id": "LR_MATH_052", "domain": "Procedural
   Reasoning", "question_text_english": "The sum of two
   numbers is 30, and their difference is 6. What are the
   two numbers?", "gold_standard_answer": "18 and 12", "
   answer_format_regex": "18 and 12|12 and 18", "
   gold_standard_reasoning": "Step 1: Let the numbers be x
    and y. x + y = 30 and x - y = 6. Step 2: Add the two
   equations. 2x = 36, so x = 18. Step 3: Substitute x
   back into the first equation. 18 + y = 30, so y = 12.
   The numbers are 18 and 12."}
{"question_id": "LR_MATH_053", "domain": "Procedural
   Reasoning", "question_text_english": "If it takes 3
   painters 4 days to paint a house, how many days would
   it take 2 painters?", "gold_standard_answer": "6", "
   answer_format_regex": "\\b6\\b", "
   gold_standard_reasoning": "Step 1: Calculate the total
   painter-days required. 3 painters * 4 days = 12 painter
   -days. Step 2: Calculate the time for the new team. 12
   painter-days / 2 painters = 6 days."}
{"question_id": "LR_MATH_054", "domain": "Procedural
   Reasoning", "question_text_english": "A bag contains 20
    apples. 1/4 are green and the rest are red. How many
   red apples are there?", "gold_standard_answer": "15", "
   answer_format_regex": "\\b15\\b", "
   gold_standard_reasoning": "Step 1: Calculate the number
    of green apples. 20 apples * (1/4) = 5 green apples.
   Step 2: Calculate the number of red apples. 20 total
   apples - 5 green apples = 15 red apples."}
{"question_id": "LR_MATH_055", "domain": "Procedural
   Reasoning", "question_text_english": "A plant grows 2
   cm every week. If it is 10 cm tall now, how tall will
   it be in 6 weeks?", "gold_standard_answer": "22", "
   answer_format_regex": "\\b22\\b", "
   gold_standard_reasoning": "Step 1: Calculate the total
   growth. 2 cm/week * 6 weeks = 12 cm. Step 2: Add the
   growth to the current height. 10 cm + 12 cm = 22 cm."}
```

{"question_id": "LR_MATH_056", "domain": "Procedural Reasoning", "question_text_english": "A rectangle has a perimeter of 30 cm. If its length is 10 cm, what is its width?", "gold_standard_answer": "5", "answer_format_regex": "\\b5\\b", "gold_standard_reasoning": "Step 1: The formula for perimeter is P = 2l + 2w. We know P=30 and l=10. Step 2: Substitute the values. 30 = 2*10 + 2w. Step 3: Solve for w. 30 = 20 + 2w. 10 = 2w. w = 5 cm."}
{"question_id": "LR_MATH_057", "domain": "Procedural Reasoning", "question_text_english": "What is the next number in the sequence: 3, 7, 11, 15, ...?", "gold_standard_answer": "19", "answer_format_regex": "\\b19\\b", "gold_standard_reasoning": "Step 1: Identify the pattern in the sequence. The difference between consecutive numbers is 4 (7-3=4, 11-7=4). Step 2: Add 4 to the last number to find the next one. 15 + 4 = 19."}
{"question_id": "LR_MATH_058", "domain": "Procedural Reasoning", "question_text_english": "A meeting is scheduled for 1.5 hours. If it starts 15 minutes late, and finishes on time, how long was the actual meeting in minutes?", "gold_standard_answer": "75", "answer_format_regex": "\\b75\\b", "gold_standard_reasoning": "Step 1: Convert the scheduled time to minutes. 1.5 hours * 60 minutes/hour = 90 minutes. Step 2: Subtract the delay from the scheduled time. 90 minutes - 15 minutes = 75 minutes."}
{"question_id": "LR_MATH_059", "domain": "Procedural Reasoning", "question_text_english": "A cake requires 200g of sugar for every 500g of flour. If you use 1 kg of flour, how much sugar do you need?", "gold_standard_answer": "400", "answer_format_regex": "\\b400\\b", "gold_standard_reasoning": "Step 1: Convert 1 kg of flour to grams. 1 kg = 1000g. Step 2: Find the ratio multiplier. 1000g / 500g = 2. Step 3: Multiply the sugar amount by the multiplier. 200g * 2 = 400g."}
{"question_id": "LR_MATH_060", "domain": "Procedural Reasoning", "question_text_english": "A store sells T-shirts for \u00a315 each. If you buy 3, you get a 10% discount on the total. How much do you pay for 3 T-shirts?", "gold_standard_answer": "40.50", "answer_format_regex": "40\\\\.50|40\\\\.5|\u00a340\\\\.50", "gold_standard_reasoning": "Step 1: Calculate the full price for 3 T-shirts. 3 * \u00a315 = \u00a345. Step 2: Calculate the discount amount. \u00a345 * 10% = \u00a34.50. Step 3: Subtract the discount from the full price. \u00a345 - \u00a34.50 = \u00a340.50."}
{"question_id": "LR_MATH_061", "domain": "Procedural Reasoning", "question_text_english": "How many sides does a hexagon have?", "gold_standard_answer": "6", "answer_format_regex": "\\b6\\b", "gold_standard_reasoning": "A hexagon is a polygon with six sides and six angles. The name is derived from the Greek words 'hex' meaning six and 'gonia' meaning corner or angle."}
{"question_id": "LR_MATH_062", "domain": "Procedural Reasoning", "question_text_english": "A flight departs

```
at 09:00 and arrives at 14:30. How long is the flight
in hours and minutes?", "gold_standard_answer": "5
hours and 30 minutes", "answer_format_regex": "5 hours
and 30 minutes", "gold_standard_reasoning": "Step 1:
Subtract the departure hour from the arrival hour. 14 -
 9 = 5 hours. Step 2: Subtract the departure minutes
from the arrival minutes. 30 - 0 = 30 minutes. Step 3:
Combine the results. The duration is 5 hours and 30
minutes."}
{"question_id": "LR_MATH_063", "domain": "Procedural
Reasoning", "question_text_english": "There are 52
cards in a standard deck. What is the probability of
drawing a King?", "gold_standard_answer": "1/13", "
answer_format_regex": "1/13", "gold_standard_reasoning
": "Step 1: A standard deck has 4 Kings. Step 2: The
total number of cards is 52. Step 3: The probability is
 the number of Kings divided by the total number of
cards, which is 4/52. Step 4: Simplify the fraction.
4/52 = 1/13."}
{"question_id": "LR_MATH_064", "domain": "Procedural
Reasoning", "question_text_english": "A school trip has
 3 buses, each carrying 45 pupils. If there are 10
teachers on the trip in total, how many people are on
the trip?", "gold_standard_answer": "145", "
answer_format_regex": "\\b145\\b", "
gold_standard_reasoning": "Step 1: Calculate the total
number of pupils. 3 buses * 45 pupils/bus = 135 pupils.
 Step 2: Add the number of teachers. 135 pupils + 10
teachers = 145 people."}
{"question_id": "LR_MATH_065", "domain": "Procedural
Reasoning", "question_text_english": "A garden is in
the shape of a right-angled triangle with base 10m and
height 6m. What is the area?", "gold_standard_answer":
"30", "answer_format_regex": "\\b30\\b", "
gold_standard_reasoning": "Step 1: The formula for the
area of a triangle is A = (1/2) * base * height. Step
2: Substitute the values. A = (1/2) * 10 * 6. Step 3:
Calculate the result. A = 30 square metres."}
{"question_id": "LR_MATH_066", "domain": "Procedural
Reasoning", "question_text_english": "A company has
1,500 employees. 40% work in sales. How many employees
do not work in sales?", "gold_standard_answer": "900",
"answer_format_regex": "\\b900\\b", "
gold_standard_reasoning": "Step 1: The percentage of
employees not in sales is 100% - 40% = 60%. Step 2:
Calculate the number of employees not in sales. 1500 *
60% = 900 employees."}
{"question_id": "LR_MATH_067", "domain": "Procedural
Reasoning", "question_text_english": "A barrel contains
 120 litres of water. If 1/3 is used for gardening and
1/4 of the remainder is used for cleaning, how much
water is left in litres?", "gold_standard_answer":
"60", "answer_format_regex": "\\b60\\b", "
gold_standard_reasoning": "Step 1: Calculate water used
 for gardening. 120 * (1/3) = 40 litres. Step 2:
Calculate remaining water. 120 - 40 = 80 litres. Step
3: Calculate water used for cleaning. 80 * (1/4) = 20
litres. Step 4: Calculate water left. 80 - 20 = 60
litres."}
```

{"question_id": "LR_MATH_068", "domain": "Procedural Reasoning", "question_text_english": "A box of 12 pencils costs \u00a33.00. How much would 30 pencils cost?", "gold_standard_answer": "7.50", "answer_format_regex": "7\\\\.50|7\\\\.5|\u00a37\\\\.50", "gold_standard_reasoning": "Step 1: Find the cost of one pencil. \u00a33.00 / 12 = \u00a30.25 per pencil. Step 2: Calculate the cost for 30 pencils. 30 * \u00a30.25 = \u00a37.50."}
{"question_id": "LR_MATH_069", "domain": "Procedural Reasoning", "question_text_english": "If you roll a standard six-sided die, what is the probability of rolling an even number?", "gold_standard_answer": "0.5", "answer_format_regex": "0\\\\.5|1/2|50%", "gold_standard_reasoning": "Step 1: The possible outcomes are 1, 2, 3, 4, 5, 6 (6 total outcomes). Step 2: The even number outcomes are 2, 4, 6 (3 favourable outcomes). Step 3: The probability is (Favourable Outcomes) / (Total Outcomes) = 3 / 6 = 1/2 or 0.5."}
{"question_id": "LR_MATH_070", "domain": "Procedural Reasoning", "question_text_english": "A recipe for 4 people requires 600g of potatoes. How many grams of potatoes are needed for 6 people?", "gold_standard_answer": "900", "answer_format_regex": "\\b900\\b", "gold_standard_reasoning": "Step 1: Calculate potatoes needed per person. 600g / 4 people = 150g per person. Step 2: Calculate total potatoes needed for 6 people. 6 people * 150g/person = 900g."}
{"question_id": "LR_MATH_071", "domain": "Procedural Reasoning", "question_text_english": "An office has 25 desks. If 60% of them are occupied, how many desks are empty?", "gold_standard_answer": "10", "answer_format_regex": "\\b10\\b", "gold_standard_reasoning": "Step 1: The percentage of empty desks is 100% - 60% = 40%. Step 2: Calculate the number of empty desks. 25 desks * 40% = 10 desks."}
{"question_id": "LR_MATH_072", "domain": "Procedural Reasoning", "question_text_english": "A car's value depreciates by 20% in its first year. If it was bought for \u00a318,000, what is its value after one year?", "gold_standard_answer": "14400", "answer_format_regex": "\\b14400\\b|\u00a314,400", "gold_standard_reasoning": "Step 1: Calculate the depreciation amount. \u00a318,000 * 20% = \u00a33,600. Step 2: Subtract the depreciation from the original value. \u00a318,000 - \u00a33,600 = \u00a314,400."}
{"question_id": "LR_MATH_073", "domain": "Procedural Reasoning", "question_text_english": "What are the prime factors of 30?", "gold_standard_answer": "2, 3, and 5", "answer_format_regex": "2, 3, and 5", "gold_standard_reasoning": "Step 1: Start dividing 30 by the smallest prime number, 2. 30 / 2 = 15. So 2 is a factor. Step 2: Now divide 15 by the next smallest prime number, 3. 15 / 3 = 5. So 3 is a factor. Step 3: 5 is a prime number. So the prime factors are 2, 3, and 5."}
{"question_id": "LR_MATH_074", "domain": "Procedural Reasoning", "question_text_english": "A train ticket costs \u00a345 for a single journey. A return ticket

```
costs \u00a370. How much do you save by buying a return
 ticket instead of two singles?", "gold_standard_answer
": "20", "answer_format_regex": "\\b20\\b|\u00a320", "
gold_standard_reasoning": "Step 1: Calculate the cost
of two single tickets. 2 * \u00a345 = \u00a390. Step 2:
 Calculate the saving. \u00a390 - \u00a370 = \u00a320
."}
{"question_id": "LR_MATH_075", "domain": "Procedural
Reasoning", "question_text_english": "A class of 24
pupils is divided into groups of 3. How many groups are
 there?", "gold_standard_answer": "8", "
answer_format_regex": "\\b8\\b", "
gold_standard_reasoning": "Step 1: Divide the total
number of pupils by the group size. 24 pupils / 3
pupils per group = 8 groups."}
{"question_id": "LR_MATH_076", "domain": "Procedural
Reasoning", "question_text_english": "A phone plan
costs \u00a320 per month plus 10 pence per text. How
much is the bill for a month with 50 texts?", "
gold_standard_answer": "25", "answer_format_regex": "\\
b25\\b|\u00a325", "gold_standard_reasoning": "Step 1:
Calculate the cost of the texts. 50 texts * 10 pence/
text = 500 pence. Step 2: Convert the text cost to
pounds. 500 pence = \u00a35. Step 3: Add the monthly
cost. \u00a320 + \u00a35 = \u00a325."}
{"question_id": "LR_MATH_077", "domain": "Procedural
Reasoning", "question_text_english": "A square has a
perimeter of 48 cm. What is its area?", "
gold_standard_answer": "144", "answer_format_regex":
"\\b144\\b", "gold_standard_reasoning": "Step 1:
Calculate the length of one side. A square has 4 equal
sides, so 48 cm / 4 = 12 cm per side. Step 2: Calculate
 the area. Area = side * side = 12 cm * 12 cm = 144
square cm."}
{"question_id": "LR_MATH_078", "domain": "Procedural
Reasoning", "question_text_english": "The average of
four numbers is 15. If three of the numbers are 10, 12,
 and 18, what is the fourth number?", "
gold_standard_answer": "20", "answer_format_regex": "\\
b20\\b", "gold_standard_reasoning": "Step 1: Calculate
the total sum of the four numbers. 15 (average) * 4 (
count) = 60. Step 2: Calculate the sum of the three
known numbers. 10 + 12 + 18 = 40. Step 3: Find the
fourth number by subtracting the sum of the three from
the total sum. 60 - 40 = 20."}
{"question_id": "LR_MATH_079", "domain": "Procedural
Reasoning", "question_text_english": "A baker uses 50g
of yeast for every 2kg of flour. What is the ratio of
yeast to flour in its simplest form?", "
gold_standard_answer": "1:40", "answer_format_regex":
"1:40", "gold_standard_reasoning": "Step 1: Convert all
 units to be the same (grams). 2kg = 2000g. Step 2:
Write the ratio. 50g : 2000g. Step 3: Simplify the
ratio by dividing both sides by the greatest common
divisor, which is 50. 50/50 : 2000/50 = 1:40."}
{"question_id": "LR_MATH_080", "domain": "Procedural
Reasoning", "question_text_english": "A tank is 3/4
full. After using 100 litres of water, it is 1/2 full.
What is the total capacity of the tank in litres?", "
```

gold_standard_answer": "400", "answer_format_regex": "\\b400\\b", "gold_standard_reasoning": "Step 1: Calculate the fraction of the tank that was used. 3/4 - 1/2 = 3/4 - 2/4 = 1/4. Step 2: We know that 1/4 of the tank's capacity is equal to 100 litres. Step 3: Calculate the total capacity. If 1/4 is 100 litres, then the full capacity (4/4) is 4 * 100 = 400 litres."}
{"question_id": "LR_MATH_081", "domain": "Procedural Reasoning", "question_text_english": "What is the area of a circle with a diameter of 20 cm? (Use pi \u2248 3.14)", "gold_standard_answer": "314", "answer_format_regex": "\\b314\\b", "gold_standard_reasoning": "Step 1: Calculate the radius from the diameter. Radius = Diameter / 2 = 20 cm / 2 = 10 cm. Step 2: Use the formula for the area of a circle: A = pi * r^2. Step 3: Substitute the values. A = 3.14 * (10^2) = 3.14 * 100 = 314 square cm."}
{"question_id": "LR_MATH_082", "domain": "Procedural Reasoning", "question_text_english": "If a car travels 180 miles in 3 hours, what is its average speed in miles per hour?", "gold_standard_answer": "60", "answer_format_regex": "\\b60\\b", "gold_standard_reasoning": "Step 1: Use the formula Speed = Distance / Time. Step 2: Substitute the values. Speed = 180 miles / 3 hours. Step 3: Calculate the result. 60 miles per hour."}
{"question_id": "LR_MATH_083", "domain": "Procedural Reasoning", "question_text_english": "A school has 600 pupils. 55% are girls. How many boys are there?", "gold_standard_answer": "270", "answer_format_regex": "\\b270\\b", "gold_standard_reasoning": "Step 1: Calculate the percentage of boys. 100% - 55% = 45%. Step 2: Calculate the number of boys. 600 pupils * 45% = 270 boys."}
{"question_id": "LR_MATH_084", "domain": "Procedural Reasoning", "question_text_english": "A bag of flour weighs 1.5 kg. If 600g is used for a recipe, how many grams of flour are left?", "gold_standard_answer": "900", "answer_format_regex": "\\b900\\b", "gold_standard_reasoning": "Step 1: Convert the initial weight to grams. 1.5 kg = 1500g. Step 2: Subtract the amount used. 1500g - 600g = 900g."}
{"question_id": "LR_MATH_085", "domain": "Procedural Reasoning", "question_text_english": "A TV is priced at \u00a3800. A shop offers a choice of a \u00a3100 discount or a 15% discount. Which is the better deal and by how much?", "gold_standard_answer": "The 15% discount is better by \u00a320", "answer_format_regex": "15% discount is better by \u00a320", "gold_standard_reasoning": "Step 1: The first deal is a \u00a3100 discount. Step 2: Calculate the second discount. \u00a3800 * 15% = \u00a3120. Step 3: Compare the discounts. \u00a3120 is greater than \u00a3100. Step 4: Calculate the difference. \u00a3120 - \u00a3100 = \u00a320. The 15% discount is better by \u00a320."}
{"question_id": "LR_MATH_086", "domain": "Procedural Reasoning", "question_text_english": "There are 25 chocolates in a box. 8 are milk chocolate, 12 are dark chocolate, and the rest are white chocolate. What

```
percentage are white chocolate?", "gold_standard_answer
": "20", "answer_format_regex": "20|20%", "
gold_standard_reasoning": "Step 1: Calculate the number
 of white chocolates. 25 - 8 - 12 = 5 white chocolates.
 Step 2: Calculate the percentage. (5 / 25) * 100 =
20%."}
{"question_id": "LR_MATH_087", "domain": "Procedural
Reasoning", "question_text_english": "A jug contains 2
litres of juice. If you pour out five 250ml glasses,
how much juice is left in the jug in millilitres?", "
gold_standard_answer": "750", "answer_format_regex":
"\\b750\\b", "gold_standard_reasoning": "Step 1:
Calculate the total juice poured out. 5 glasses * 250ml
/glass = 1250ml. Step 2: Convert the initial volume to
millilitres. 2 litres = 2000ml. Step 3: Calculate the
remaining juice. 2000ml - 1250ml = 750ml."}
{"question_id": "LR_MATH_088", "domain": "Procedural
Reasoning", "question_text_english": "What is the
greatest common divisor of 24 and 36?", "
gold_standard_answer": "12", "answer_format_regex": "\\
b12\\b", "gold_standard_reasoning": "Step 1: Find the
factors of 24: 1, 2, 3, 4, 6, 8, 12, 24. Step 2: Find
the factors of 36: 1, 2, 3, 4, 6, 9, 12, 18, 36. Step
3: Identify the common factors: 1, 2, 3, 4, 6, 12. Step
 4: The greatest of these is 12."}
{"question_id": "LR_MATH_089", "domain": "Procedural
Reasoning", "question_text_english": "A company's stock
 price was \u00a350. It went up by 10% on Monday and
then down by 10% on Tuesday. What was the final price
?", "gold_standard_answer": "49.50", "
answer_format_regex": "49\\\\.50|49\\\\.5|\u00a349
\\\\.50", "gold_standard_reasoning": "Step 1: Calculate
 the price after Monday's increase. \u00a350 * 1.10 = \
u00a355. Step 2: Calculate the price after Tuesday's
decrease. \u00a355 * 0.90 = \u00a349.50."}
{"question_id": "LR_MATH_090", "domain": "Procedural
Reasoning", "question_text_english": "The sum of the
angles in a triangle is how many degrees?", "
gold_standard_answer": "180", "answer_format_regex":
"\\b180\\b", "gold_standard_reasoning": "The sum of the
 interior angles of any triangle is always 180 degrees.
 This is a fundamental theorem of Euclidean geometry."}
{"question_id": "LR_MATH_091", "domain": "Procedural
Reasoning", "question_text_english": "A box of 8
cupcakes costs \u00a312. A box of 12 cupcakes costs \
u00a315. Which is the better value for money?", "
gold_standard_answer": "The box of 12", "
answer_format_regex": "box of 12", "
gold_standard_reasoning": "Step 1: Calculate the price
per cupcake for the first box. \u00a312 / 8 = \u00a31
.50 per cupcake. Step 2: Calculate the price per
cupcake for the second box. \u00a315 / 12 = \u00a31.25
per cupcake. Step 3: Compare the prices. \u00a31.25 is
cheaper than \u00a31.50, so the box of 12 is better
value."}
{"question_id": "LR_MATH_092", "domain": "Procedural
Reasoning", "question_text_english": "A train is
scheduled to arrive at 17:10. It is delayed by 45
minutes. What is the new arrival time?", "
```

gold_standard_answer": "17:55", "answer_format_regex": "17:55", "gold_standard_reasoning": "Step 1: Add the delay to the scheduled arrival time. 17:10 + 45 minutes. Step 2: Add the minutes. 10 + 45 = 55 minutes. The new arrival time is 17:55."}
{"question_id": "LR_MATH_093", "domain": "Procedural Reasoning", "question_text_english": "A library charges a late fee of 25 pence per day. If a book is 2 weeks late, how much is the fine in pounds?", "gold_standard_answer": "3.50", "answer_format_regex": "3\\\\.50|3\\\\.5|\u00a33\\\\.50", "gold_standard_reasoning": "Step 1: Convert the late period to days. 2 weeks * 7 days/week = 14 days. Step 2: Calculate the total fine in pence. 14 days * 25 pence/day = 350 pence. Step 3: Convert the fine to pounds. 350 pence = \u00a33.50."}
{"question_id": "LR_MATH_094", "domain": "Procedural Reasoning", "question_text_english": "What is the next prime number after 19?", "gold_standard_answer": "23", "answer_format_regex": "\\b23\\b", "gold_standard_reasoning": "A prime number is a number greater than 1 that has no positive divisors other than 1 and itself. The number 20 is divisible by 2, 21 by 3, 22 by 2. The number 23 is only divisible by 1 and 23, so it is the next prime number."}
{"question_id": "LR_MATH_095", "domain": "Procedural Reasoning", "question_text_english": "A recipe needs 3 eggs to make 12 muffins. How many eggs are needed to make 20 muffins?", "gold_standard_answer": "5", "answer_format_regex": "\\b5\\b", "gold_standard_reasoning": "Step 1: Find how many muffins one egg makes. 12 muffins / 3 eggs = 4 muffins per egg. Step 2: Calculate how many eggs are needed for 20 muffins. 20 muffins / 4 muffins per egg = 5 eggs."}
{"question_id": "LR_MATH_096", "domain": "Procedural Reasoning", "question_text_english": "An equilateral triangle has a perimeter of 24 cm. What is the length of one side?", "gold_standard_answer": "8", "answer_format_regex": "\\b8\\b", "gold_standard_reasoning": "An equilateral triangle has three equal sides. To find the length of one side, divide the perimeter by 3. 24 cm / 3 = 8 cm."}
{"question_id": "LR_MATH_097", "domain": "Procedural Reasoning", "question_text_english": "A job is advertised with a salary of \u00a360,000 per year. What is the monthly salary?", "gold_standard_answer": "5000", "answer_format_regex": "\\b5000\\b|\u00a35,000", "gold_standard_reasoning": "There are 12 months in a year. To find the monthly salary, divide the annual salary by 12. \u00a360,000 / 12 = \u00a35,000."}
{"question_id": "LR_MATH_098", "domain": "Procedural Reasoning", "question_text_english": "A bottle of water costs \u00a30.80. A pack of 6 costs \u00a34.50. How much cheaper is each bottle when bought in a pack?", "gold_standard_answer": "0.05", "answer_format_regex": "0\\\\.05|5 pence|\u00a30\\\\.05", "gold_standard_reasoning": "Step 1: Calculate the price per bottle in the pack. \u00a34.50 / 6 = \u00a30.75 per

```
 bottle. Step 2: Calculate the difference in price. \
u00a30.80 - \u00a30.75 = \u00a30.05, or 5 pence."}
{"question_id": "LR_MATH_099", "domain": "Procedural
Reasoning", "question_text_english": "There are 1,000
metres in a kilometre. How many metres are there in 3.5
 kilometres?", "gold_standard_answer": "3500", "
answer_format_regex": "\\b3500\\b", "
gold_standard_reasoning": "To convert kilometres to
metres, you multiply by 1,000. 3.5 km * 1000 m/km =
3500 metres."}
{"question_id": "LR_MATH_100", "domain": "Procedural
Reasoning", "question_text_english": "A film rental
costs \u00a34 for 3 nights. If you keep it for a week
(7 nights), and each extra night costs \u00a31.50, what
 is the total cost?", "gold_standard_answer": "10.00",
"answer_format_regex": "10\\\\.00|\\\\b10\\\\b|\u00a310
", "gold_standard_reasoning": "Step 1: Calculate the
number of extra nights. 7 nights - 3 nights = 4 extra
nights. Step 2: Calculate the cost of the extra nights.
 4 * \u00a31.50 = \u00a36.00. Step 3: Calculate the
total cost. \u00a34.00 (initial cost) + \u00a36.00 (
extra cost) = \u00a310.00."}
{"question_id": "LR_MATH_101", "domain": "Procedural
Reasoning", "question_text_english": "A garden hose
fills a 200-litre water butt in 40 minutes. What is the
 flow rate in litres per minute?", "
gold_standard_answer": "5", "answer_format_regex": "\\
b5\\b", "gold_standard_reasoning": "Flow rate is volume
 divided by time. 200 litres / 40 minutes = 5 litres
per minute."}
{"question_id": "LR_MATH_102", "domain": "Procedural
Reasoning", "question_text_english": "Three friends
share a pizza. Tom eats 1/3, Jane eats 1/4. What
fraction of the pizza is left for Chris?", "
gold_standard_answer": "5/12", "answer_format_regex":
"5/12", "gold_standard_reasoning": "Step 1: Find a
common denominator for 3 and 4, which is 12. Tom eats
4/12 and Jane eats 3/12. Step 2: Calculate the total
eaten. 4/12 + 3/12 = 7/12. Step 3: Subtract the eaten
part from the whole (12/12). 12/12 - 7/12 = 5/12."}
{"question_id": "LR_MATH_103", "domain": "Procedural
Reasoning", "question_text_english": "A book costs \
u00a315. It is on sale with a 30% discount. What is the
 sale price?", "gold_standard_answer": "10.50", "
answer_format_regex": "10\\\\.50|10\\\\.5|\u00a310
\\\\.50", "gold_standard_reasoning": "Step 1: Calculate
 the discount amount. \u00a315 * 30% = \u00a34.50. Step
 2: Subtract the discount from the original price. \
u00a315 - \u00a34.50 = \u00a310.50."}
{"question_id": "LR_MATH_104", "domain": "Procedural
Reasoning", "question_text_english": "A car travels 300
 kilometres on 25 litres of petrol. How many kilometres
 per litre does the car get?", "gold_standard_answer":
"12", "answer_format_regex": "\\b12\\b", "
gold_standard_reasoning": "To find the efficiency,
divide the distance by the fuel used. 300 kilometres /
25 litres = 12 kilometres per litre."}
{"question_id": "LR_MATH_105", "domain": "Procedural
Reasoning", "question_text_english": "If an investment
```

```
of \u00a32,000 earns 5% simple interest per year, how
much interest is earned in 3 years?", "
gold_standard_answer": "300", "answer_format_regex":
"\\b300\\b|\u00a3300", "gold_standard_reasoning": "Step
 1: Calculate the interest per year. \u00a32,000 * 5% =
 \u00a3100 per year. Step 2: Calculate the total
interest for 3 years. \u00a3100 * 3 = \u00a3300."}
{"question_id": "LR_MATH_106", "domain": "Procedural
Reasoning", "question_text_english": "A room measures 4
 metres by 5 metres. What is the area of the floor in
square metres?", "gold_standard_answer": "20", "
answer_format_regex": "\\b20\\b", "
gold_standard_reasoning": "The area of a rectangle is
length times width. 4 metres * 5 metres = 20 square
metres."}
{"question_id": "LR_MATH_107", "domain": "Procedural
Reasoning", "question_text_english": "A box contains 60
 sweets. 1/3 are red, 1/4 are green, and the rest are
yellow. How many yellow sweets are there?", "
gold_standard_answer": "25", "answer_format_regex": "\\
b25\\b", "gold_standard_reasoning": "Step 1: Calculate
the number of red sweets. 60 * (1/3) = 20. Step 2:
Calculate the number of green sweets. 60 * (1/4) = 15.
Step 3: Calculate the number of yellow sweets. 60 - 20
- 15 = 25."}
{"question_id": "LR_MATH_108", "domain": "Procedural
Reasoning", "question_text_english": "What is 15% of
200?", "gold_standard_answer": "30", "
answer_format_regex": "\\b30\\b", "
gold_standard_reasoning": "To calculate the percentage,
 multiply the number by the percentage as a decimal.
200 * 0.15 = 30."}
{"question_id": "LR_MATH_109", "domain": "Procedural
Reasoning", "question_text_english": "If you cycle at
12 km/h for 2.5 hours, how far have you cycled?", "
gold_standard_answer": "30", "answer_format_regex": "\\
b30\\b", "gold_standard_reasoning": "Distance is speed
multiplied by time. 12 km/h * 2.5 hours = 30 km."}
{"question_id": "LR_MATH_110", "domain": "Procedural
Reasoning", "question_text_english": "A packet of
biscuits contains 20 biscuits. A family eats 4 biscuits
 a day. How many days will the packet last?", "
gold_standard_answer": "5", "answer_format_regex": "\\
b5\\b", "gold_standard_reasoning": "To find how long it
 will last, divide the total number of biscuits by the
number eaten per day. 20 biscuits / 4 biscuits per day
= 5 days."}
{"question_id": "LR_MATH_111", "domain": "Procedural
Reasoning", "question_text_english": "A meal for two
people costs \u00a340. They want to leave a 10% tip.
How much is the tip?", "gold_standard_answer": "4", "
answer_format_regex": "\\b4\\b|\u00a34", "
gold_standard_reasoning": "To calculate the tip,
multiply the cost by the percentage. \u00a340 * 10% = \
u00a34."}
{"question_id": "LR_MATH_112", "domain": "Procedural
Reasoning", "question_text_english": "The temperature
drops from 8 degrees Celsius to -3 degrees Celsius. How
 many degrees did it fall?", "gold_standard_answer":
```

```
   "11", "answer_format_regex": "\\b11\\b", "
   gold_standard_reasoning": "The total fall is the
   difference between the starting and ending temperatures
   . 8 - (-3) = 8 + 3 = 11 degrees."}
{"question_id": "LR_MATH_113", "domain": "Procedural
   Reasoning", "question_text_english": "A work shift is 8
    hours long, including a 45-minute unpaid break. How
   many minutes do you get paid for?", "
   gold_standard_answer": "435", "answer_format_regex":
   "\\b435\\b", "gold_standard_reasoning": "Step 1:
   Convert the total shift time to minutes. 8 hours * 60
   minutes/hour = 480 minutes. Step 2: Subtract the unpaid
    break time. 480 minutes - 45 minutes = 435 minutes."}
{"question_id": "LR_MATH_114", "domain": "Procedural
   Reasoning", "question_text_english": "A pack of 4
   batteries costs \u00a32.80. What is the cost per
   battery?", "gold_standard_answer": "0.70", "
   answer_format_regex": "0\\\\.70|0\\\\.7|70p", "
   gold_standard_reasoning": "To find the cost per battery
   , divide the total cost by the number of batteries. \
   u00a32.80 / 4 = \u00a30.70, or 70 pence."}
{"question_id": "LR_MATH_115", "domain": "Procedural
   Reasoning", "question_text_english": "A square has a
   perimeter of 100 cm. What is its area?", "
   gold_standard_answer": "625", "answer_format_regex":
   "\\b625\\b", "gold_standard_reasoning": "Step 1: Find
   the length of one side. 100 cm / 4 = 25 cm. Step 2:
   Calculate the area. Area = side * side = 25 cm * 25 cm
   = 625 square cm."}
{"question_id": "LR_MATH_116", "domain": "Procedural
   Reasoning", "question_text_english": "A recipe requires
    300ml of milk. If you only have a 1-litre carton, what
    percentage of the milk will you use?", "
   gold_standard_answer": "30", "answer_format_regex":
   "30|30%", "gold_standard_reasoning": "Step 1: Convert
   the carton size to ml. 1 litre = 1000ml. Step 2:
   Calculate the percentage used. (300ml / 1000ml) * 100 =
    30%."}
{"question_id": "LR_MATH_117", "domain": "Procedural
   Reasoning", "question_text_english": "There are 7 days
   in a week. How many full weeks are there in a year of
   365 days?", "gold_standard_answer": "52", "
   answer_format_regex": "\\b52\\b", "
   gold_standard_reasoning": "To find the number of weeks,
    divide the total days by 7. 365 / 7 = 52 with a
   remainder of 1. So there are 52 full weeks."}
{"question_id": "LR_MATH_118", "domain": "Procedural
   Reasoning", "question_text_english": "If you score 18
   out of 20 on a test, what is your score as a percentage
   ?", "gold_standard_answer": "90", "answer_format_regex
   ": "90|90%", "gold_standard_reasoning": "To find the
   percentage, divide your score by the total and multiply
    by 100. (18 / 20) * 100 = 0.9 * 100 = 90%."}
{"question_id": "LR_MATH_119", "domain": "Procedural
   Reasoning", "question_text_english": "A bag contains
   only red and blue marbles. The ratio of red to blue is
   3:5. If there are 15 red marbles, how many blue marbles
    are there?", "gold_standard_answer": "25", "
   answer_format_regex": "\\b25\\b", "
```

```
   gold_standard_reasoning": "Step 1: Find the multiplier
   for the ratio. 15 red marbles / 3 parts = 5. Step 2:
   Apply the multiplier to the blue marbles part of the
   ratio. 5 parts * 5 = 25 blue marbles."}
{"question_id": "LR_MATH_120", "domain": "Procedural
   Reasoning", "question_text_english": "A train journey
   takes 4 hours and 15 minutes. If it arrives at 13:00,
   what time did it depart?", "gold_standard_answer":
   "08:45", "answer_format_regex": "08:45|8:45 AM", "
   gold_standard_reasoning": "Step 1: Subtract the hours
   from the arrival time. 13:00 - 4 hours = 09:00. Step 2:
    Subtract the minutes from the result. 09:00 - 15
   minutes = 08:45."}
{"question_id": "LR_MATH_121", "domain": "Procedural
   Reasoning", "question_text_english": "A piece of string
    is 5 metres long. You cut off 80 centimetres. How much
    string is left in metres?", "gold_standard_answer":
   "4.2", "answer_format_regex": "4\\\\.2", "
   gold_standard_reasoning": "Step 1: Convert centimetres
   to metres. 80 cm = 0.8 m. Step 2: Subtract the cut
   length from the total length. 5 m - 0.8 m = 4.2 metres
   ."}
{"question_id": "LR_MATH_122", "domain": "Procedural
   Reasoning", "question_text_english": "A company has a
   budget of \u00a310,000. It spends \u00a34,500 on
   salaries and \u00a32,000 on rent. What percentage of
   the budget is left?", "gold_standard_answer": "35", "
   answer_format_regex": "35|35%", "
   gold_standard_reasoning": "Step 1: Calculate total
   spending. \u00a34,500 + \u00a32,000 = \u00a36,500. Step
    2: Calculate the remaining budget. \u00a310,000 - \
   u00a36,500 = \u00a33,500. Step 3: Calculate the
   remaining percentage. (\u00a33,500 / \u00a310,000) *
   100 = 35%."}
{"question_id": "LR_MATH_123", "domain": "Procedural
   Reasoning", "question_text_english": "The sum of the
   ages of a mother and daughter is 50. The mother is 24
   years older than the daughter. How old is the daughter
   ?", "gold_standard_answer": "13", "answer_format_regex
   ": "\\b13\\b", "gold_standard_reasoning": "Step 1: Let
   the daughter's age be 'd'. The mother's age is 'd +
   24'. Step 2: Set up the equation. d + (d + 24) = 50.
   Step 3: Solve for d. 2d + 24 = 50. 2d = 26. d = 13. The
    daughter is 13 years old."}
{"question_id": "LR_MATH_124", "domain": "Procedural
   Reasoning", "question_text_english": "A car park
   charges \u00a32 for the first hour and \u00a31.50 for
   each additional hour. How much does it cost to park for
    4 hours?", "gold_standard_answer": "6.50", "
   answer_format_regex": "6\\\\.50|6\\\\.5|\u00a36
   \\\\.50", "gold_standard_reasoning": "Step 1: The first
    hour costs \u00a32. Step 2: There are 3 additional
   hours. Step 3: Calculate the cost of additional hours.
   3 * \u00a31.50 = \u00a34.50. Step 4: Calculate the
   total cost. \u00a32 + \u00a34.50 = \u00a36.50."}
{"question_id": "LR_MATH_125", "domain": "Procedural
   Reasoning", "question_text_english": "What is the
   average of the numbers 10, 20, 30, and 40?", "
   gold_standard_answer": "25", "answer_format_regex": "\\
```

```
   b25\\b", "gold_standard_reasoning": "Step 1: Sum the
   numbers. 10 + 20 + 30 + 40 = 100. Step 2: Divide the
   sum by the count of the numbers. 100 / 4 = 25."}
{"question_id": "LR_MATH_126", "domain": "Procedural
   Reasoning", "question_text_english": "A shop sells 120
   newspapers a day. 65% are sold in the morning. How many
    newspapers are sold in the afternoon?", "
   gold_standard_answer": "42", "answer_format_regex": "\\
   b42\\b", "gold_standard_reasoning": "Step 1: Calculate
   the percentage sold in the afternoon. 100% - 65% = 35%.
    Step 2: Calculate the number of newspapers sold in the
    afternoon. 120 * 35% = 42."}
{"question_id": "LR_MATH_127", "domain": "Procedural
   Reasoning", "question_text_english": "A tank can hold
   5000 litres of oil. If it is 85% full, how many litres
   of oil are in the tank?", "gold_standard_answer":
   "4250", "answer_format_regex": "\\b4250\\b", "
   gold_standard_reasoning": "To find the amount of oil,
   multiply the total capacity by the percentage full.
   5000 litres * 85% = 4250 litres."}
{"question_id": "LR_MATH_128", "domain": "Procedural
   Reasoning", "question_text_english": "A runner's
   average speed is 8 mph. How long will it take them to
   run a marathon of 26 miles?", "gold_standard_answer":
   "3.25 hours", "answer_format_regex": "3\\\\.25", "
   gold_standard_reasoning": "Time equals distance divided
    by speed. 26 miles / 8 mph = 3.25 hours."}
{"question_id": "LR_MATH_129", "domain": "Procedural
   Reasoning", "question_text_english": "If an item costs
   \u00a340 and the price is increased by 5%, what is the
   new price?", "gold_standard_answer": "42", "
   answer_format_regex": "\\b42\\b|\u00a342", "
   gold_standard_reasoning": "Step 1: Calculate the price
   increase. \u00a340 * 5% = \u00a32. Step 2: Add the
   increase to the original price. \u00a340 + \u00a32 = \
   u00a342."}
{"question_id": "LR_MATH_130", "domain": "Procedural
   Reasoning", "question_text_english": "A rectangle's
   length is twice its width. If the width is 6 cm, what
   is the area of the rectangle?", "gold_standard_answer":
    "72", "answer_format_regex": "\\b72\\b", "
   gold_standard_reasoning": "Step 1: Calculate the length
   . 6 cm (width) * 2 = 12 cm. Step 2: Calculate the area.
    Area = length * width = 12 cm * 6 cm = 72 square cm."}
{"question_id": "LR_MATH_131", "domain": "Procedural
   Reasoning", "question_text_english": "You have a 5-
   litre jug and a 3-litre jug. How can you measure
   exactly 4 litres of water?", "gold_standard_answer": "
   Fill 5L, pour into 3L, empty 3L, pour remaining 2L from
    5L into 3L, fill 5L, pour from 5L to fill 3L", "
   answer_format_regex": "4 litres", "
   gold_standard_reasoning": "1. Fill the 5-litre jug. 2.
   Pour from the 5L jug to fill the 3L jug. You now have 2
   L in the 5L jug. 3. Empty the 3L jug. 4. Pour the 2L
   from the 5L jug into the 3L jug. 5. Fill the 5L jug
   again. 6. Pour from the 5L jug to fill the 3L jug (
   which already has 2L in it), this will use 1L. 7. The 5
   L jug now contains exactly 4 litres."}
```

{"question_id": "LR_MATH_132", "domain": "Procedural Reasoning", "question_text_english": "A cinema has 20 rows of seats. Each row has 25 seats. If 420 seats are sold for a film, how many are empty?", "gold_standard_answer": "80", "answer_format_regex": "\\b80\\b", "gold_standard_reasoning": "Step 1: Calculate the total number of seats. 20 rows * 25 seats/row = 500 seats. Step 2: Calculate the number of empty seats. 500 total seats - 420 sold seats = 80 empty seats."}
{"question_id": "LR_MATH_133", "domain": "Procedural Reasoning", "question_text_english": "What is the simple interest on \u00a3500 at a rate of 4% per annum for 5 years?", "gold_standard_answer": "100", "answer_format_regex": "\\b100\\b|\u00a3100", "gold_standard_reasoning": "Simple interest is calculated as Principal * Rate * Time. \u00a3500 * 4% * 5 years = \u00a3500 * 0.04 * 5 = \u00a3100."}
{"question_id": "LR_MATH_134", "domain": "Procedural Reasoning", "question_text_english": "If a dozen eggs cost \u00a32.40, how much does one egg cost?", "gold_standard_answer": "0.20", "answer_format_regex": "0\\\\.20|20p|\u00a30\\\\.20", "gold_standard_reasoning": "A dozen is 12. To find the cost of one egg, divide the total cost by 12. \u00a32.40 / 12 = \u00a30.20, or 20 pence."}
{"question_id": "LR_MATH_135", "domain": "Procedural Reasoning", "question_text_english": "A recipe needs 150g of butter. If you are making a double batch, how many grams of butter do you need?", "gold_standard_answer": "300", "answer_format_regex": "\\b300\\b", "gold_standard_reasoning": "To make a double batch, you need twice the amount of each ingredient. 150g * 2 = 300g."}
{"question_id": "LR_MATH_136", "domain": "Procedural Reasoning", "question_text_english": "The sum of the angles in a quadrilateral is how many degrees?", "gold_standard_answer": "360", "answer_format_regex": "\\b360\\b", "gold_standard_reasoning": "The sum of the interior angles of any simple quadrilateral is 360 degrees. This can be shown by dividing the quadrilateral into two triangles, each of which has angles summing to 180 degrees."}
{"question_id": "LR_MATH_137", "domain": "Procedural Reasoning", "question_text_english": "A train travels at 120 km/h. How far does it travel in 20 minutes?", "gold_standard_answer": "40", "answer_format_regex": "\\b40\\b", "gold_standard_reasoning": "Step 1: Convert time to hours. 20 minutes is 1/3 of an hour. Step 2: Calculate distance. Distance = Speed * Time = 120 km/h * (1/3) h = 40 km."}
{"question_id": "LR_MATH_138", "domain": "Procedural Reasoning", "question_text_english": "A shop's weekly profit is \u00a35,200. What is its average daily profit?", "gold_standard_answer": "742.86", "answer_format_regex": "742\\\\.86", "gold_standard_reasoning": "To find the average daily profit, divide the weekly profit by 7. \u00a35200 / 7 = \u00a3742.857... which rounds to \u00a3742.86."}

{"question_id": "LR_MATH_139", "domain": "Procedural Reasoning", "question_text_english": "If a shirt is discounted by 25% to a sale price of \u00a330, what was its original price?", "gold_standard_answer": "40", "answer_format_regex": "\\b40\\b|\u00a340", "gold_standard_reasoning": "Step 1: The sale price represents 100% - 25% = 75% of the original price. Step 2: If \u00a330 is 75%, then 1% is \u00a330 / 75 = \u00a30.40. Step 3: The original price (100%) is 100 * \u00a30.40 = \u00a340."}
{"question_id": "LR_MATH_140", "domain": "Procedural Reasoning", "question_text_english": "A box of tiles contains 20 tiles. Each tile is a square with a side of 10 cm. What is the total area the tiles can cover in square metres?", "gold_standard_answer": "0.2", "answer_format_regex": "0\\\\.2", "gold_standard_reasoning": "Step 1: Calculate the area of one tile in square cm. 10 cm * 10 cm = 100 sq cm. Step 2: Calculate the total area in square cm. 20 tiles * 100 sq cm/tile = 2000 sq cm. Step 3: Convert square cm to square metres (10,000 sq cm = 1 sq m). 2000 / 10000 = 0.2 square metres."}
{"question_id": "LR_MATH_141", "domain": "Procedural Reasoning", "question_text_english": "A cyclist's average speed is 18 km/h. How many metres do they travel per minute?", "gold_standard_answer": "300", "answer_format_regex": "\\b300\\b", "gold_standard_reasoning": "Step 1: Convert speed to metres per hour. 18 km/h = 18,000 m/h. Step 2: Convert speed to metres per minute. 18,000 m/h / 60 min/h = 300 metres per minute."}
{"question_id": "LR_MATH_142", "domain": "Procedural Reasoning", "question_text_english": "A bag contains 50 balls. 30% are red, 20% are blue, and the rest are green. How many green balls are there?", "gold_standard_answer": "25", "answer_format_regex": "\\b25\\b", "gold_standard_reasoning": "Step 1: Calculate the percentage of green balls. 100% - 30% - 20% = 50%. Step 2: Calculate the number of green balls. 50 balls * 50% = 25 green balls."}
{"question_id": "LR_MATH_143", "domain": "Procedural Reasoning", "question_text_english": "A phone costs \u00a3600. You pay a 25% deposit. How much is the remaining balance?", "gold_standard_answer": "450", "answer_format_regex": "\\b450\\b|\u00a3450", "gold_standard_reasoning": "Step 1: The remaining balance is 100% - 25% = 75% of the cost. Step 2: Calculate the remaining balance. \u00a3600 * 75% = \u00a3450."}
{"question_id": "LR_MATH_144", "domain": "Procedural Reasoning", "question_text_english": "If it is 16:00 in London (GMT), what time is it in New York (GMT-5)?", "gold_standard_answer": "11:00", "answer_format_regex": "11:00|11 AM", "gold_standard_reasoning": "New York is 5 hours behind London. To find the time, subtract 5 hours from London time. 16:00 - 5 hours = 11:00."}
{"question_id": "LR_MATH_145", "domain": "Procedural Reasoning", "question_text_english": "A book has an ISBN number. What does the acronym ISBN stand for?", "

gold_standard_answer": "International Standard Book Number", "answer_format_regex": "International Standard Book Number", "gold_standard_reasoning": "ISBN stands for International Standard Book Number. It is a unique numeric commercial book identifier."}
{"question_id": "LR_MATH_146", "domain": "Procedural Reasoning", "question_text_english": "A bill for a meal is \u00a372. Three friends agree to split it equally. How much does each person pay?", "gold_standard_answer": "24", "answer_format_regex": "\\b24\\b|\u00a324", "gold_standard_reasoning": "To split the bill equally, divide the total cost by the number of people. \u00a372 / 3 = \u00a324."}
{"question_id": "LR_MATH_147", "domain": "Procedural Reasoning", "question_text_english": "A car uses 8 litres of petrol per 100 kilometres. How many litres will it use on a 350-kilometre journey?", "gold_standard_answer": "28", "answer_format_regex": "\\b28\\b", "gold_standard_reasoning": "Step 1: Find the multiplier for the distance. 350 km / 100 km = 3.5. Step 2: Multiply the fuel consumption by this multiplier. 8 litres * 3.5 = 28 litres."}
{"question_id": "LR_MATH_148", "domain": "Procedural Reasoning", "question_text_english": "A test starts at 10:30 and lasts for 75 minutes. At what time does it end?", "gold_standard_answer": "11:45", "answer_format_regex": "11:45", "gold_standard_reasoning": "75 minutes is 1 hour and 15 minutes. Add this to the start time. 10:30 + 1 hour = 11:30. 11:30 + 15 minutes = 11:45."}
{"question_id": "LR_MATH_149", "domain": "Procedural Reasoning", "question_text_english": "A cube has a volume of 27 cubic cm. What is the length of one side?", "gold_standard_answer": "3", "answer_format_regex": "\\b3\\b", "gold_standard_reasoning": "The volume of a cube is the side length cubed. To find the side length, take the cube root of the volume. The cube root of 27 is 3. So the side is 3 cm."}
{"question_id": "LR_MATH_150", "domain": "Procedural Reasoning", "question_text_english": "In a group of 80 people, 48 have brown hair. What percentage of the group does not have brown hair?", "gold_standard_answer": "40", "answer_format_regex": "40|40%", "gold_standard_reasoning": "Step 1: Find the number of people who do not have brown hair. 80 - 48 = 32. Step 2: Calculate the percentage. (32 / 80) * 100 = 40%."}
{"question_id": "LR_MATH_151", "domain": "Procedural Reasoning", "question_text_english": "A company produces 500 widgets per day. How many widgets will it produce in a 5-day work week?", "gold_standard_answer": "2500", "answer_format_regex": "\\b2500\\b", "gold_standard_reasoning": "To find the total production, multiply the daily production by the number of days. 500 widgets/day * 5 days = 2500 widgets."}
{"question_id": "LR_MATH_152", "domain": "Procedural Reasoning", "question_text_english": "If a map scale is 1cm to 5km, how many cm on the map represents a real distance of 35km?", "gold_standard_answer": "7", "answer_format_regex": "\\b7\\b", "

gold_standard_reasoning": "To find the map distance, divide the real distance by the scale factor. 35 km / 5 km per cm = 7 cm."}
{"question_id": "LR_MATH_153", "domain": "Procedural Reasoning", "question_text_english": "A library has 12,000 books. 1/5 are fiction. Of the remaining books, 1/4 are history books. How many history books are there?", "gold_standard_answer": "2400", "answer_format_regex": "\\b2400\\b", "gold_standard_reasoning": "Step 1: Calculate the number of fiction books. 12,000 * (1/5) = 2,400. Step 2: Calculate the remaining non-fiction books. 12,000 - 2,400 = 9,600. Step 3: Calculate the number of history books. 9,600 * (1/4) = 2,400."}
{"question_id": "LR_MATH_154", "domain": "Procedural Reasoning", "question_text_english": "A recipe needs 2 cups of flour for every 1 cup of sugar. If you use 3 cups of sugar, how many cups of flour do you need?", "gold_standard_answer": "6", "answer_format_regex": "\\b6\\b", "gold_standard_reasoning": "The ratio of flour to sugar is 2:1. If you use 3 cups of sugar, you need to multiply the flour amount by 3. 2 cups of flour * 3 = 6 cups of flour."}
{"question_id": "LR_MATH_155", "domain": "Procedural Reasoning", "question_text_english": "A worker earns \u00a312 per hour. How much do they earn in an 8-hour day?", "gold_standard_answer": "96", "answer_format_regex": "\\b96\\b|\u00a396", "gold_standard_reasoning": "To find the total earnings, multiply the hourly rate by the number of hours worked. \u00a312/hour * 8 hours = \u00a396."}
{"question_id": "LR_MATH_156", "domain": "Procedural Reasoning", "question_text_english": "A water tank is 50% full. If 200 litres are added, it becomes 75% full. What is the total capacity of the tank?", "gold_standard_answer": "800", "answer_format_regex": "\\b800\\b", "gold_standard_reasoning": "Step 1: The 200 litres added represents a 75% - 50% = 25% increase in fullness. Step 2: If 25% (or 1/4) of the tank is 200 litres, the total capacity is 4 * 200 = 800 litres."}
{"question_id": "LR_MATH_157", "domain": "Procedural Reasoning", "question_text_english": "A shirt costs \u00a325. The shop adds a 20% VAT. What is the final price for the customer?", "gold_standard_answer": "30", "answer_format_regex": "\\b30\\b|\u00a330", "gold_standard_reasoning": "Step 1: Calculate the VAT amount. \u00a325 * 20% = \u00a35. Step 2: Add the VAT to the original price. \u00a325 + \u00a35 = \u00a330."}
{"question_id": "LR_MATH_158", "domain": "Procedural Reasoning", "question_text_english": "If you walk 1.5 km in 20 minutes, what is your speed in km/h?", "gold_standard_answer": "4.5", "answer_format_regex": "4\\\\.5", "gold_standard_reasoning": "Step 1: There are three 20-minute periods in an hour. Step 2: To find your speed per hour, multiply the distance walked by 3. 1.5 km * 3 = 4.5 km/h."}
{"question_id": "LR_MATH_159", "domain": "Procedural Reasoning", "question_text_english": "The product of two numbers is 48 and their sum is 14. What are the two

```
 numbers?", "gold_standard_answer": "6 and 8", "
answer_format_regex": "6 and 8|8 and 6", "
gold_standard_reasoning": "We need two numbers that
multiply to 48 and add to 14. The pairs of factors for
48 are (1,48), (2,24), (3,16), (4,12), and (6,8). The
pair that sums to 14 is 6 and 8."}
{"question_id": "LR_MATH_160", "domain": "Procedural
Reasoning", "question_text_english": "A train leaves at
 22:30 on Monday and arrives at 05:45 on Tuesday. What
is the duration of the journey?", "gold_standard_answer
": "7 hours and 15 minutes", "answer_format_regex": "7
hours and 15 minutes", "gold_standard_reasoning": "Step
 1: Calculate the time from 22:30 to midnight. This is
1 hour and 30 minutes. Step 2: Add the time from
midnight to 05:45. This is 5 hours and 45 minutes. Step
 3: Sum the two durations. 1h 30m + 5h 45m = 6h 75m,
which is 7 hours and 15 minutes."}
{"question_id": "LR_MATH_161", "domain": "Procedural
Reasoning", "question_text_english": "What is the area
of a triangle with a base of 12 cm and a height of 8 cm
?", "gold_standard_answer": "48", "answer_format_regex
": "\\b48\\b", "gold_standard_reasoning": "The area of
a triangle is (1/2) * base * height. (1/2) * 12 cm * 8
cm = 48 square cm."}
{"question_id": "LR_MATH_162", "domain": "Procedural
Reasoning", "question_text_english": "A company's
shares are worth \u00a31.50 each. If you buy 500 shares
, how much do you pay?", "gold_standard_answer": "750",
 "answer_format_regex": "\\b750\\b|\u00a3750", "
gold_standard_reasoning": "To find the total cost,
multiply the number of shares by the price per share.
500 * \u00a31.50 = \u00a3750."}
{"question_id": "LR_MATH_163", "domain": "Procedural
Reasoning", "question_text_english": "In a survey of
200 people, 120 said they prefer tea over coffee. What
percentage of people prefer tea?", "
gold_standard_answer": "60", "answer_format_regex":
"60|60%", "gold_standard_reasoning": "To find the
percentage, divide the number of tea drinkers by the
total number of people and multiply by 100. (120 / 200)
 * 100 = 0.6 * 100 = 60%."}
{"question_id": "LR_MATH_164", "domain": "Procedural
Reasoning", "question_text_english": "A car travels 30
miles on one gallon of petrol. How many gallons are
needed for a 210-mile trip?", "gold_standard_answer":
"7", "answer_format_regex": "\\b7\\b", "
gold_standard_reasoning": "To find the gallons needed,
divide the total distance by the car's efficiency. 210
miles / 30 miles per gallon = 7 gallons."}
{"question_id": "LR_MATH_165", "domain": "Procedural
Reasoning", "question_text_english": "A rectangle has
an area of 96 square cm. If its width is 8 cm, what is
its length?", "gold_standard_answer": "12", "
answer_format_regex": "\\b12\\b", "
gold_standard_reasoning": "The area of a rectangle is
length times width. To find the length, divide the area
 by the width. 96 sq cm / 8 cm = 12 cm."}
{"question_id": "LR_MATH_166", "domain": "Procedural
Reasoning", "question_text_english": "If 5 miles is
```

approximately 8 kilometres, how many kilometres is 30 miles?", "gold_standard_answer": "48", "answer_format_regex": "\\b48\\b", "gold_standard_reasoning": "Step 1: Find the multiplier. 30 miles / 5 miles = 6. Step 2: Multiply the kilometre equivalent by the multiplier. 8 km * 6 = 48 km."}
{"question_id": "LR_MATH_167", "domain": "Procedural Reasoning", "question_text_english": "A pack of 500 sheets of paper is 5 cm thick. What is the thickness of a single sheet of paper in millimetres?", "gold_standard_answer": "0.1", "answer_format_regex": "0\\\\.1", "gold_standard_reasoning": "Step 1: Calculate the thickness of one sheet in cm. 5 cm / 500 sheets = 0.01 cm. Step 2: Convert cm to mm. 0.01 cm * 10 mm/cm = 0.1 mm."}
{"question_id": "LR_MATH_168", "domain": "Procedural Reasoning", "question_text_english": "A baker can make 4 cakes in an hour. How many hours will it take to make 18 cakes?", "gold_standard_answer": "4.5", "answer_format_regex": "4\\\\.5", "gold_standard_reasoning": "To find the time needed, divide the total number of cakes by the rate of production. 18 cakes / 4 cakes per hour = 4.5 hours."}
{"question_id": "LR_MATH_169", "domain": "Procedural Reasoning", "question_text_english": "A television originally priced at \u00a3400 is sold for \u00a3320. What was the percentage discount?", "gold_standard_answer": "20", "answer_format_regex": "20|20%", "gold_standard_reasoning": "Step 1: Find the discount amount. \u00a3400 - \u00a3320 = \u00a380. Step 2: Calculate the percentage discount. (\u00a380 / \u00a3400) * 100 = 20%."}
{"question_id": "LR_MATH_170", "domain": "Procedural Reasoning", "question_text_english": "What is the smallest number that is divisible by 2, 3, and 5?", "gold_standard_answer": "30", "answer_format_regex": "\\b30\\b", "gold_standard_reasoning": "This is the lowest common multiple (LCM) of 2, 3, and 5. Since they are all prime numbers, their LCM is their product. 2 * 3 * 5 = 30."}
{"question_id": "LR_MATH_171", "domain": "Procedural Reasoning", "question_text_english": "A farmer has 180 sheep. The ratio of male to female sheep is 1:5. How many male sheep are there?", "gold_standard_answer": "30", "answer_format_regex": "\\b30\\b", "gold_standard_reasoning": "Step 1: The total number of parts in the ratio is 1 + 5 = 6. Step 2: Find the value of one part. 180 sheep / 6 parts = 30 sheep per part. Step 3: There is 1 part male sheep, so there are 30 male sheep."}
{"question_id": "LR_MATH_172", "domain": "Procedural Reasoning", "question_text_english": "A phone has 64GB of storage. If 48GB is used, what percentage of storage is free?", "gold_standard_answer": "25", "answer_format_regex": "25|25%", "gold_standard_reasoning": "Step 1: Find the amount of free storage. 64GB - 48GB = 16GB. Step 2: Calculate the percentage free. (16GB / 64GB) * 100 = 25%."}

{"question_id": "LR_MATH_173", "domain": "Procedural Reasoning", "question_text_english": "A box of 6 eggs costs \u00a31.50. A box of 12 eggs costs \u00a32.40. How much cheaper per egg is the larger box?", "gold_standard_answer": "0.05", "answer_format_regex": "0\\\\.05|5p|\u00a30\\\\.05", "gold_standard_reasoning": "Step 1: Price per egg in the small box. \u00a31.50 / 6 = \u00a30.25. Step 2: Price per egg in the large box. \u00a32.40 / 12 = \u00a30.20. Step 3: Find the difference. \u00a30.25 - \u00a30.20 = \u00a30.05, or 5 pence."}
{"question_id": "LR_MATH_174", "domain": "Procedural Reasoning", "question_text_english": "A car is travelling at 90 km/h. How many metres does it travel in one second?", "gold_standard_answer": "25", "answer_format_regex": "\\b25\\b", "gold_standard_reasoning": "Step 1: Convert km/h to m/h. 90 km/h = 90,000 m/h. Step 2: Convert m/h to m/s. 90,000 m/h / 3600 s/h = 25 metres per second."}
{"question_id": "LR_MATH_175", "domain": "Procedural Reasoning", "question_text_english": "What is the perimeter of a regular octagon with a side length of 6 cm?", "gold_standard_answer": "48", "answer_format_regex": "\\b48\\b", "gold_standard_reasoning": "A regular octagon has 8 equal sides. The perimeter is the side length multiplied by the number of sides. 6 cm * 8 = 48 cm."}
{"question_id": "LR_MATH_176", "domain": "Procedural Reasoning", "question_text_english": "A book has 400 pages. On Monday you read 10% of the book. On Tuesday you read 25% of the remaining pages. How many pages have you read in total?", "gold_standard_answer": "130", "answer_format_regex": "\\b130\\b", "gold_standard_reasoning": "Step 1: Pages read on Monday. 400 * 10% = 40 pages. Step 2: Remaining pages. 400 - 40 = 360 pages. Step 3: Pages read on Tuesday. 360 * 25% = 90 pages. Step 4: Total pages read. 40 + 90 = 130 pages."}
{"question_id": "LR_MATH_177", "domain": "Procedural Reasoning", "question_text_english": "A restaurant bill is \u00a350. You add a 12.5% service charge. What is the total bill?", "gold_standard_answer": "56.25", "answer_format_regex": "56\\\\.25|\u00a356\\\\.25", "gold_standard_reasoning": "Step 1: Calculate the service charge amount. \u00a350 * 12.5% = \u00a36.25. Step 2: Add the service charge to the bill. \u00a350 + \u00a36.25 = \u00a356.25."}
{"question_id": "LR_MATH_178", "domain": "Procedural Reasoning", "question_text_english": "A recipe needs 400g of flour to serve 4 people. You are cooking for 5 people. How much flour do you need?", "gold_standard_answer": "500", "answer_format_regex": "\\b500\\b", "gold_standard_reasoning": "Step 1: Find the flour needed per person. 400g / 4 people = 100g per person. Step 2: Calculate the flour for 5 people. 5 people * 100g/person = 500g."}
{"question_id": "LR_MATH_179", "domain": "Procedural Reasoning", "question_text_english": "If you invest \u00a31000 for 2 years with a compound interest rate of

```
5% per year, what is the total amount at the end of the
 2 years?", "gold_standard_answer": "1102.50", "
answer_format_regex": "1102\\\\.50|\u00a31102\\\\.50",
"gold_standard_reasoning": "Step 1: After year 1, the
amount is \u00a31000 * 1.05 = \u00a31050. Step 2: After
 year 2, the amount is \u00a31050 * 1.05 = \u00a31102
.50."}
{"question_id": "LR_MATH_180", "domain": "Procedural
Reasoning", "question_text_english": "A car's fuel
efficiency is 50 miles per gallon. How many gallons
would be needed for a 450-mile trip?", "
gold_standard_answer": "9", "answer_format_regex": "\\
b9\\b", "gold_standard_reasoning": "To find the fuel
needed, divide the distance by the fuel efficiency. 450
 miles / 50 miles per gallon = 9 gallons."}
{"question_id": "LR_MATH_181", "domain": "Procedural
Reasoning", "question_text_english": "The angles in a
triangle are in the ratio 1:2:3. What is the size of
the largest angle?", "gold_standard_answer": "90", "
answer_format_regex": "\\b90\\b", "
gold_standard_reasoning": "Step 1: The sum of angles in
 a triangle is 180 degrees. The total parts in the
ratio is 1+2+3=6. Step 2: The value of one part is 180
/ 6 = 30 degrees. Step 3: The largest angle is 3 parts,
 so 3 * 30 = 90 degrees."}
{"question_id": "LR_MATH_182", "domain": "Procedural
Reasoning", "question_text_english": "A shop buys a
coat for \u00a350 and sells it for \u00a380. What is
the percentage profit mark-up?", "gold_standard_answer
": "60", "answer_format_regex": "60|60%", "
gold_standard_reasoning": "Step 1: Calculate the profit
 amount. \u00a380 - \u00a350 = \u00a330. Step 2:
Calculate the percentage mark-up based on the cost
price. (\u00a330 / \u00a350) * 100 = 60%."}
{"question_id": "LR_MATH_183", "domain": "Procedural
Reasoning", "question_text_english": "A playlist
contains 120 songs. The average song length is 3
minutes. How long is the entire playlist in hours?", "
gold_standard_answer": "6", "answer_format_regex": "\\
b6\\b", "gold_standard_reasoning": "Step 1: Calculate
the total length in minutes. 120 songs * 3 minutes/song
 = 360 minutes. Step 2: Convert minutes to hours. 360
minutes / 60 minutes/hour = 6 hours."}
{"question_id": "LR_MATH_184", "domain": "Procedural
Reasoning", "question_text_english": "If 8 out of 40
students in a class wear glasses, what percentage of
students do not wear glasses?", "gold_standard_answer":
 "80", "answer_format_regex": "80|80%", "
gold_standard_reasoning": "Step 1: Find the number of
students who do not wear glasses. 40 - 8 = 32. Step 2:
Calculate the percentage. (32 / 40) * 100 = 80%."}
{"question_id": "LR_MATH_185", "domain": "Procedural
Reasoning", "question_text_english": "A window is 1.2
metres wide and 1.5 metres high. What is its area in
square metres?", "gold_standard_answer": "1.8", "
answer_format_regex": "1\\\\.8", "
gold_standard_reasoning": "The area is width multiplied
 by height. 1.2 metres * 1.5 metres = 1.8 square metres
."}
```

{"question_id": "LR_MATH_186", "domain": "Procedural Reasoning", "question_text_english": "What is the lowest common multiple of 8 and 12?", "gold_standard_answer": "24", "answer_format_regex": "\\b24\\b", "gold_standard_reasoning": "Multiples of 8 are 8, 16, 24, 32... Multiples of 12 are 12, 24, 36... The lowest number that appears in both lists is 24."}
{"question_id": "LR_MATH_187", "domain": "Procedural Reasoning", "question_text_english": "A train has 512 passengers. At the first stop, half of the passengers get off. At the second stop, half of the remaining passengers get off. How many passengers are left?", "gold_standard_answer": "128", "answer_format_regex": "\\b128\\b", "gold_standard_reasoning": "Step 1: After the first stop, 512 / 2 = 256 passengers remain. Step 2: After the second stop, 256 / 2 = 128 passengers are left."}
{"question_id": "LR_MATH_188", "domain": "Procedural Reasoning", "question_text_english": "A runner aims to run 30 miles in a week. By Wednesday, she has run 12 miles. What percentage of her weekly goal has she completed?", "gold_standard_answer": "40", "answer_format_regex": "40|40%", "gold_standard_reasoning": "To find the percentage completed, divide the miles run by the total goal and multiply by 100. (12 / 30) * 100 = 0.4 * 100 = 40%."}
{"question_id": "LR_MATH_189", "domain": "Procedural Reasoning", "question_text_english": "A bag costs \u00a390. In a sale, it is reduced by 1/3. What is the sale price?", "gold_standard_answer": "60", "answer_format_regex": "\\b60\\b|\u00a360", "gold_standard_reasoning": "Step 1: Calculate the discount amount. \u00a390 * (1/3) = \u00a330. Step 2: Subtract the discount from the original price. \u00a390 - \u00a330 = \u00a360."}
{"question_id": "LR_MATH_190", "domain": "Procedural Reasoning", "question_text_english": "A cyclist travels 45 km in 3 hours. What is their average speed in km/h?", "gold_standard_answer": "15", "answer_format_regex": "\\b15\\b", "gold_standard_reasoning": "Average speed is total distance divided by total time. 45 km / 3 hours = 15 km/h."}
{"question_id": "LR_MATH_191", "domain": "Procedural Reasoning", "question_text_english": "A box contains blue and red pens. There are 36 pens in total. If 25% are blue, how many are red?", "gold_standard_answer": "27", "answer_format_regex": "\\b27\\b", "gold_standard_reasoning": "Step 1: Calculate the number of blue pens. 36 * 25% = 9 blue pens. Step 2: Calculate the number of red pens. 36 - 9 = 27 red pens."}
{"question_id": "LR_MATH_192", "domain": "Procedural Reasoning", "question_text_english": "A room requires 24 square metres of carpet. If the carpet costs \u00a315 per square metre, what is the total cost?", "gold_standard_answer": "360", "answer_format_regex": "\\b360\\b|\u00a3360", "gold_standard_reasoning": "To find the total cost, multiply the area by the cost per square metre. 24 * \u00a315 = \u00a3360."}

```
{"question_id": "LR_MATH_193", "domain": "Procedural
   Reasoning", "question_text_english": "If an item costs
   \u00a380 including 20% VAT, what is the price before
   VAT was added?", "gold_standard_answer": "66.67", "
   answer_format_regex": "66\\\\.67", "
   gold_standard_reasoning": "Step 1: The price including
   VAT represents 120% of the original price. Step 2: To
   find the original price, divide the final price by
   1.20. \u00a380 / 1.20 = \u00a366.67."}
{"question_id": "LR_MATH_194", "domain": "Procedural
   Reasoning", "question_text_english": "What is the value
    of 2 to the power of 6?", "gold_standard_answer":
   "64", "answer_format_regex": "\\b64\\b", "
   gold_standard_reasoning": "2 to the power of 6 means
   multiplying 2 by itself 6 times. 2 * 2 * 2 * 2 * 2 * 2
   = 64."}
{"question_id": "LR_MATH_195", "domain": "Procedural
   Reasoning", "question_text_english": "A car is bought
   for \u00a310,000 and sold 3 years later for \u00a34
   ,000. What was the total depreciation?", "
   gold_standard_answer": "6000", "answer_format_regex":
   "\\b6000\\b|\u00a36,000", "gold_standard_reasoning": "
   Depreciation is the difference between the purchase
   price and the selling price. \u00a310,000 - \u00a34,000
    = \u00a36,000."}
{"question_id": "LR_MATH_196", "domain": "Procedural
   Reasoning", "question_text_english": "A recipe needs
   250g of flour, 150g of sugar, and 100g of butter. What
   is the ratio of flour to sugar to butter in its
   simplest form?", "gold_standard_answer": "5:3:2", "
   answer_format_regex": "5:3:2", "gold_standard_reasoning
   ": "The ratio is 250:150:100. To simplify, find the
   greatest common divisor, which is 50. Divide each part
   of the ratio by 50. 250/50 : 150/50 : 100/50 = 5:3:2."}
{"question_id": "LR_MATH_197", "domain": "Procedural
   Reasoning", "question_text_english": "A phone battery
   lasts for 10 hours of use. If you have used it for 4.5
   hours, what percentage of the battery is remaining?", "
   gold_standard_answer": "55", "answer_format_regex":
   "55|55%", "gold_standard_reasoning": "Step 1: Find the
   remaining time. 10 hours - 4.5 hours = 5.5 hours. Step
   2: Calculate the remaining percentage. (5.5 / 10) * 100
    = 55%."}
{"question_id": "LR_MATH_198", "domain": "Procedural
   Reasoning", "question_text_english": "The sum of two
   consecutive numbers is 31. What are the numbers?", "
   gold_standard_answer": "15 and 16", "
   answer_format_regex": "15 and 16|16 and 15", "
   gold_standard_reasoning": "Let the numbers be x and x
   +1. The equation is x + (x+1) = 31. 2x + 1 = 31. 2x =
   30. x = 15. The numbers are 15 and 16."}
{"question_id": "LR_MATH_199", "domain": "Procedural
   Reasoning", "question_text_english": "A journey of 150
   miles takes 2.5 hours. What was the average speed in
   mph?", "gold_standard_answer": "60", "
   answer_format_regex": "\\b60\\b", "
   gold_standard_reasoning": "Average speed is distance
   divided by time. 150 miles / 2.5 hours = 60 mph."}
```

```
{"question_id": "LR_MATH_200", "domain": "Procedural
   Reasoning", "question_text_english": "A company has 80
   employees. 12 of them are left-handed. What percentage
   of employees are right-handed?", "gold_standard_answer
   ": "85", "answer_format_regex": "85|85%", "
   gold_standard_reasoning": "Step 1: Find the number of
   right-handed employees. 80 - 12 = 68. Step 2: Calculate
    the percentage. (68 / 80) * 100 = 0.85 * 100 = 85%."}
```


C Data pipeline program files

The following are the program listings used. They are presented in the order that they are to be run in.

build_corpus.py

```
"""
build_corpus.py

Purpose:
Reads the human-editable master corpus from a TSV file and
    converts it into
a syntactically perfect, machine-readable JSON Lines (
   JSONL) file. This script
is the first step in the data pipeline, ensuring the
   source data is clean
and correctly formatted for all subsequent processing.

Authors:
Hejroe, Gemini

Version:
1.0

Last Updated:
16 November 2025

Inputs:
- ../data/master_corpus.tsv: A Tab-Separated Values file
   containing the full
question corpus with headers.

Outputs:
- ../data/questions_en_uk.jsonl: The master English (UK)
   question corpus in
JSONL format, with all strings and special characters
   correctly escaped.

License:
MIT License
-------------------------------------------------------------------------

"""

import csv
import json
from pathlib import Path

# --- Configuration ---
```

```
# Use pathlib to create robust, cross-platform paths
    relative to this script's location.
# Path(__file__).resolve() -> gets the full path to this
    script
# .parent -> gets the 'src' directory
# .parent -> gets the root project directory (Multi-Ling-
    Rubric/)
BASE_DIR = Path(__file__).resolve().parent.parent

SOURCE_TSV_FILE = BASE_DIR / 'data' / 'master_corpus.tsv'
OUTPUT_JSONL_FILE = BASE_DIR / 'data' / 'questions_en_uk.
    jsonl'

def main():
"""
Main function to read the TSV and generate the JSONL file.
"""
corpus = []
print(f"Reading master corpus from: {SOURCE_TSV_FILE}")
try:
with open(SOURCE_TSV_FILE, 'r', encoding='utf-8') as f:
# Use csv.DictReader with 'excel-tab' dialect for TSV
    files.
reader = csv.DictReader(f, dialect='excel-tab')
for row in reader:
# If the reasoning field is empty in the TSV, it should be
     None.
if not row.get('gold_standard_reasoning'):
row['gold_standard_reasoning'] = None
corpus.append(row)

except FileNotFoundError:
print(f"Error: Master corpus file not found at '{
    SOURCE_TSV_FILE}'.")
return
except Exception as e:
print(f"An error occurred while reading the TSV file: {e
    }")
return

# Ensure the output directory exists
OUTPUT_JSONL_FILE.parent.mkdir(parents=True, exist_ok=True
    )

# Write the data to the JSONL file
print(f"Generating JSONL file at: {OUTPUT_JSONL_FILE}")
try:
with open(OUTPUT_JSONL_FILE, 'w', encoding='utf-8') as f:
for entry in corpus:
# Exclude keys with None values for a cleaner output file.
clean_entry = {k: v for k, v in entry.items() if v is not
    None}

# json.dumps handles all necessary escaping automatically
    and correctly.
json_line = json.dumps(clean_entry)
f.write(json_line + '\n')
except Exception as e:
```

```
print(f"An error occurred while writing the JSONL file: {e
    }")
return

print(f"Successfully generated '{OUTPUT_JSONL_FILE.name}'
    with {len(corpus)} questions.")

if __name__ == '__main__':
main()
```

translate_corpus.py

```
"""
translate_corpus.py

Purpose:
Takes the master English JSONL corpus and translates the
    questions into the
target languages (DE, ES). It applies a 'Round-Trip
    Translation' quality
gate, using semantic similarity to ensure a high degree of
     translational
fidelity.

Authors:
Hejroe, Gemini

Version:
1.0

Last Updated:
16 November 2025

Pre-requisites:
- An internet connect is required for the translation.

Inputs:
- ../data/questions_en_uk.jsonl: The master English (UK)
    corpus.

Outputs:
- ../translation_outputs/questions_[lang]_[timestamp].
    jsonl: Validated,
translated question files for each target language.
- ../translation_outputs/translation_log_[timestamp].txt:
    A simple log file
detailing the Pass/Fail status for each question's
    translation.

License:
MIT License
------------------------------------------------------------------------

"""

import json
from deep_translator import GoogleTranslator
from sentence_transformers import SentenceTransformer,
    util
from tqdm import tqdm
```

```
from datetime import datetime, timezone
from pathlib import Path

# --- Configuration ---
BASE_DIR = Path(__file__).resolve().parent.parent
SOURCE_FILE = BASE_DIR / 'data' / 'questions_en_uk.jsonl'
OUTPUT_DIR = BASE_DIR / 'translation_outputs'
TARGET_LANGUAGES = ['de', 'es']
SIMILARITY_THRESHOLD = 0.95

# --- File Naming with Timestamp ---
TIMESTAMP_UTC = datetime.now(timezone.utc).strftime('%Y%m%
    d_%H%M%S')
LOG_FILE = OUTPUT_DIR / f'translation_log_{TIMESTAMP_UTC}.
    txt'

# --- Load Models ---
print("Loading semantic similarity model (this may take a
    moment on first run)...")
try:
similarity_model = SentenceTransformer('all-MiniLM-L6-v2')
print("Model loaded successfully.")
except Exception as e:
print(f"Error loading SentenceTransformer model. Ensure
    you have an internet connection.")
print(f"Error: {e}")
exit()

def translate_and_validate(question_text, target_lang):
"""
Performs a round-trip translation and validates semantic
    similarity.
Returns (translated_text, score, status).
"""
try:
to_target_translator = GoogleTranslator(source='en',
    target=target_lang)
translated_text = to_target_translator.translate(
    question_text)

from_target_translator = GoogleTranslator(source=
    target_lang, target='en')
round_trip_text = from_target_translator.translate(
    translated_text)

embedding1 = similarity_model.encode(question_text,
    convert_to_tensor=True)
embedding2 = similarity_model.encode(round_trip_text,
    convert_to_tensor=True)
cosine_score = util.pytorch_cos_sim(embedding1, embedding2
    ).item()

if cosine_score >= SIMILARITY_THRESHOLD:
return translated_text, cosine_score, "Pass"
else:
return None, cosine_score, "Fail"

except Exception as e:
return None, 0, f"Error: {e}"
```

```
def main():
"""
Main function to process the source file and generate
   translated files and a log.
"""
OUTPUT_DIR.mkdir(parents=True, exist_ok=True)

try:
with open(SOURCE_FILE, 'r', encoding='utf-8') as f:
source_questions = [json.loads(line) for line in f]
except FileNotFoundError:
print(f"Error: Source file '{SOURCE_FILE}' not found.
   Please run build_corpus.py first.")
return

# Prepare the log file header
with open(LOG_FILE, 'w', encoding='utf-8') as log_f:
log_f.write(f"Translation Validation Log - {TIMESTAMP_UTC
   }\n")
log_f.write(f"Source File: {SOURCE_FILE.name}\n")
log_f.write(f"Similarity Threshold: {SIMILARITY_THRESHOLD
   }\n")
log_f.write("="*50 + "\n")

for lang in TARGET_LANGUAGES:
print(f"\n--- Processing translations for language: {lang.
   upper()} ---")
translated_corpus = []

with open(LOG_FILE, 'a', encoding='utf-8') as log_f:
log_f.write(f"\n--- Language: {lang.upper()} ---\n")

for question in tqdm(source_questions, desc=f"Translating
   to {lang.upper()}"):
question_id = question['question_id']
original_text = question['question_text_english']

translated_text, score, status = translate_and_validate(
   original_text, lang)

log_f.write(f"{question_id}: {status}\n")

if status == "Pass":
new_question = question.copy()
new_question['question_text'] = translated_text
del new_question['question_text_english']
translated_corpus.append(new_question)

# Write the validated corpus to its timestamped file
output_file = OUTPUT_DIR / f"questions_{lang}_{
   TIMESTAMP_UTC}.jsonl"
with open(output_file, 'w', encoding='utf-8') as f:
for item in translated_corpus:
f.write(json.dumps(item) + '\n')

pass_count = len(translated_corpus)
total_count = len(source_questions)
fail_count = total_count - pass_count
```

```
print(f"Successfully created '{output_file.name}'.")
print(f"Results for {lang.upper()}: {pass_count} Passed /
    {fail_count} Failed.")

print(f"\nTranslation process complete. Log saved to '{
    LOG_FILE}'.")

if __name__ == '__main__':
main()
```

run_experiments.py

```
"""
run_experiments.py

Purpose:
Executes the core experimental runs. It systematically
    queries a list of
LLMs for every validated question in every language (EN,
    DE, ES) and saves
the complete, raw JSON responses to a timestamped results
    file.

Authors:
Hejroe, Gemini

Version:
1.0

Last Updated:
16 November 2025

Pre-requisites:
- Ollama must be running and available via the API.

Inputs:
- ../data/questions_en_uk.jsonl
- ../translation_outputs/questions_de_[latest].jsonl
- ../translation_outputs/questions_es_[latest].jsonl

Outputs:
- ../experimental_results/raw_results_[timestamp].jsonl: A
     single, large
JSONL file containing all raw model responses for the
    entire experiment.

License:
MIT License
------------------------------------------------------------------------

"""

import json
import requests
from datetime import datetime, timezone
from tqdm import tqdm
import time
import os
import glob
```

```
from pathlib import Path

# --- Configuration ---
BASE_DIR = Path(__file__).resolve().parent.parent
OLLAMA_API_ENDPOINT = 'http://localhost:11434/api/generate
    '
MODELS_TO_TEST = [
"llama3:8b", "llama3.1:8b", "llama3.2:3b", "falcon3:10b",
    "gpt-oss:20b",
"deepseek-r1:8b", "qwen3:8b", "phi4:14b", "granite3.3:8b",
"gemma3:12b", "gemma3n:e4b"
]

# --- Dynamic File Naming and Finding ---
TIMESTAMP_UTC = datetime.now(timezone.utc).strftime('%Y%m%
    d_%H%M%S')
OUTPUT_DIR = BASE_DIR / 'experimental_results'
RAW_RESULTS_FILE = OUTPUT_DIR / f'raw_results_{
    TIMESTAMP_UTC}.jsonl'

def find_latest_file(pattern):
"""Finds the most recently created file matching a pattern
    ."""
try:
list_of_files = glob.glob(pattern)
if not list_of_files: return None
return max(list_of_files, key=os.path.getctime)
except Exception as e:
print(f"Error finding file for pattern {pattern}: {e}")
return None

# Use Path objects for robust file path construction
QUESTION_FILES = {
        'EN': BASE_DIR / 'data' / 'questions_en_uk.jsonl',
        'DE': find_latest_file(str(BASE_DIR / '
            translation_outputs' / 'questions_de_*.jsonl'))
            ,
        'ES': find_latest_file(str(BASE_DIR / '
            translation_outputs' / 'questions_es_*.jsonl'))
}

def query_model(model_name, prompt):
"""Sends a prompt to a model via the Ollama API."""
payload = {"model": model_name, "prompt": prompt, "stream
    ": False}
try:
response = requests.post(OLLAMA_API_ENDPOINT, json=payload
    , timeout=120)
response.raise_for_status()
return response.json()
except requests.exceptions.RequestException as e:
return {"error": str(e)}

def main():
"""Main function to run all experiments."""
OUTPUT_DIR.mkdir(parents=True, exist_ok=True)

print("--- Starting Experiment Run ---")
```

```
print(f"Results will be saved to: {RAW_RESULTS_FILE.name
    }")

test_counter = 0
with open(RAW_RESULTS_FILE, 'w') as f: pass

for lang_code, file_path in QUESTION_FILES.items():
if file_path is None or not os.path.exists(file_path):
print(f"Warning: No question file found for language {
    lang_code}. Skipping.")
continue

print(f"\nLoading questions for language: {lang_code} from
     '{os.path.basename(file_path)}'")
with open(file_path, 'r', encoding='utf-8') as f:
questions = [json.loads(line) for line in f]

total_tests_for_lang = len(MODELS_TO_TEST) * len(questions
    )

with tqdm(total=total_tests_for_lang, desc=f"Testing
    Language: {lang_code}") as pbar:
for model in MODELS_TO_TEST:
for question in questions:
prompt = question.get('question_text', question.get('
    question_text_english'))
if not prompt:
pbar.update(1)
continue

raw_response = query_model(model, prompt)

result_record = {
        "test_id": f"run_{TIMESTAMP_UTC}_{test_counter:05d
            }",
        "question_id": question['question_id'],
        "model_identifier": model,
        "language": lang_code,
        "prompt_text": prompt,
        "raw_response": raw_response,
        "timestamp_utc": datetime.now(timezone.utc).
            isoformat()
}

with open(RAW_RESULTS_FILE, 'a', encoding='utf-8') as f:
f.write(json.dumps(result_record) + '\n')

test_counter += 1
pbar.update(1)

print(f"\n--- Experiment run complete. ---")
print(f"Saved {test_counter} results to '{RAW_RESULTS_FILE
    }'.")

if __name__ == '__main__':
main()
```

score_results.py

```
"""
```

```
score_results.py

Purpose:
Applies the final, calibrated Hybrid Automated Scoring
   protocol to the raw
experimental results. This is a fully automated script
   that categorises every
response, removing the need for subjective manual review.

Authors:
Hejroe, Gemini

Version:
1.0

Last Updated:
16 November 2025

Inputs:
- ../experimental_results/raw_results_[latest].jsonl: The
   raw experimental data.
- ../data/questions_en_uk.jsonl: The master corpus with
   answers and reasoning.

Outputs:
- ../analysis_outputs/final_scored_results_[timestamp].csv
   : The final, clean,
and fully scored dataset ready for analysis.

License:
MIT License
--------------------------------------------------------------------------

"""

import json
import pandas as pd
import re
from tqdm import tqdm
import os
import glob
from sentence_transformers import SentenceTransformer,
   util
from datetime import datetime, timezone
from pathlib import Path

# --- Configuration ---
BASE_DIR = Path(__file__).resolve().parent.parent
RAW_RESULTS_DIR = BASE_DIR / 'experimental_results'
ANALYSIS_DIR = BASE_DIR / 'analysis_outputs'
MASTER_CORPUS_FILE = BASE_DIR / 'data' / 'questions_en_uk.
   jsonl'

# --- Scoring Logic Constants ---
SCORE_CORRECT = 1.0
SCORE_CORRECT_PROCESS = 0.5
SCORE_IDK = 0.25
SCORE_AMBIGUOUS = 0.0
SCORE_INCORRECT_GUESS = -0.5
```

```
SCORE_FABRICATION = -1.0
SCORE_INCORRECT = -1.0

# --- FINAL CALIBRATED THRESHOLDS ---
REASONING_SIMILARITY_HIGH = 0.70
REASONING_SIMILARITY_LOW = 0.60

IDK_KEYWORDS = ["i don't know", "i do not know", "cannot
   answer", "unable to answer", "as an ai", "i am unable"]

# --- Load Semantic Similarity Model ---
print("Loading semantic similarity model...")
try:
similarity_model = SentenceTransformer('all-MiniLM-L6-v2')
print("Model loaded successfully.")
except Exception as e:
print(f"Error loading SentenceTransformer model: {e}")
exit()


def find_latest_file(directory, pattern):
"""Finds the most recently created file in a directory
   matching a pattern."""
try:
search_path = os.path.join(directory, pattern)
list_of_files = glob.glob(search_path)
if not list_of_files: return None
return max(list_of_files, key=os.path.getctime)
except Exception as e:
print(f"Error finding file for pattern {pattern}: {e}")
return None

def load_all_questions():
"""Loads the single, master EN(UK) question file."""
all_questions = {}
if not MASTER_CORPUS_FILE.exists():
print(f"Error: Master question file '{MASTER_CORPUS_FILE}'
    not found.")
return None

with open(MASTER_CORPUS_FILE, 'r', encoding='utf-8') as f:
for line in f:
try:
question = json.loads(line)
all_questions[question['question_id']] = question
except json.JSONDecodeError:
print(f"Warning: Skipping corrupted line in {
   MASTER_CORPUS_FILE}")
return all_questions

def score_response(row, questions_dict):
"""
Applies the hybrid automated scoring logic to a single
   result row.
Returns a tuple of (score, category, reasoning_similarity)
   .
"""
question_id = row['question_id']
question_data = questions_dict.get(question_id)
```

```
if not question_data: return 0, "QuestionDataMissing",
    None

raw_response_data = row.get('raw_response', {})
if not isinstance(raw_response_data, dict): return 0, "
    MalformedResponse", None

response_text = raw_response_data.get('response', '').
    lower()

if not response_text or "error" in raw_response_data:
    return 0, "APIError", None

# 1. Check for IDK first
if any(keyword in response_text for keyword in
    IDK_KEYWORDS):
return SCORE_IDK, "IDK", None

# 2. Check for correct final answer
regex = question_data.get('answer_format_regex', '^$')
is_answer_correct = bool(re.search(regex.lower(),
    response_text))

# --- Apply Scoring Decision Tree ---
# Factual Accuracy questions have simple scoring
if question_data.get('domain') == "Factual Accuracy":
score = SCORE_CORRECT if is_answer_correct else
    SCORE_INCORRECT
category = "Correct" if is_answer_correct else "Incorrect"
return score, category, None

# Procedural Reasoning questions use the hybrid logic
elif question_data.get('domain') == "Procedural Reasoning
    ":
gold_reasoning = question_data.get('
    gold_standard_reasoning')

if not gold_reasoning:
# If a reasoning question has no gold standard, we can't
    judge it. Score as ambiguous.
return SCORE_AMBIGUOUS, "MissingGoldReasoning", None

# Calculate reasoning similarity
try:
embedding1 = similarity_model.encode(gold_reasoning,
    convert_to_tensor=True)
embedding2 = similarity_model.encode(response_text,
    convert_to_tensor=True)
reasoning_score = util.pytorch_cos_sim(embedding1,
    embedding2).item()
except Exception:
return SCORE_AMBIGUOUS, "SimilarityError", None

# Apply the final, fully automated logic
if is_answer_correct:
if reasoning_score >= REASONING_SIMILARITY_HIGH:
return SCORE_CORRECT, "Correct", reasoning_score
elif reasoning_score < REASONING_SIMILARITY_LOW:
```

```
return SCORE_FABRICATION, "Fabrication", reasoning_score
else: # Ambiguous reasoning is a final neutral category
return SCORE_AMBIGUOUS, "AmbiguousReasoning",
    reasoning_score
else: # Answer is incorrect
if reasoning_score >= REASONING_SIMILARITY_HIGH:
return SCORE_CORRECT_PROCESS, "
    CorrectProcess_IncorrectResult", reasoning_score
elif reasoning_score < REASONING_SIMILARITY_LOW:
return SCORE_INCORRECT, "Incorrect", reasoning_score
else: # Ambiguous reasoning is a final neutral category
return SCORE_AMBIGUOUS, "AmbiguousReasoning",
    reasoning_score

# Fallback for any other case
return SCORE_INCORRECT, "UnknownDomain_Incorrect", None

def robust_read_jsonl(file_path):
"""Reads a JSON Lines file, skipping any corrupted lines
    ."""
data = []
with open(file_path, 'r', encoding='utf-8') as f:
for i, line in enumerate(f):
if not line.strip(): continue
try:
data.append(json.loads(line))
except json.JSONDecodeError:
print(f"Warning: Skipping corrupted JSON object on line {i
    +1} in {file_path}")
return pd.DataFrame(data)

def main():
"""Main function to orchestrate the scoring process."""
ANALYSIS_DIR.mkdir(parents=True, exist_ok=True)

raw_results_file = find_latest_file(RAW_RESULTS_DIR, '
    raw_results_*.jsonl')
if not raw_results_file:
print(f"Error: No raw results file found in '{
    RAW_RESULTS_DIR}'. Please run run_experiments.py first
    .")
return

print("Loading questions...")
questions_dict = load_all_questions()
if not questions_dict: return

print(f"Loading results from '{os.path.basename(
    raw_results_file)}'...")
results_df = robust_read_jsonl(raw_results_file)
if results_df.empty:
print("Error: The results file is empty or could not be
    read.")
return

print("Applying final, fully automated scoring logic...")
tqdm.pandas(desc="Scoring responses")
```

```
scored_data = results_df.progress_apply(lambda row:
    score_response(row, questions_dict), axis=1)
results_df[['score', 'score_category', '
    reasoning_similarity']] = pd.DataFrame(scored_data.
    tolist(), index=results_df.index)

results_df['domain'] = results_df['question_id'].apply(
    lambda qid: questions_dict.get(qid, {}).get('domain'))

output_columns = ['question_id', 'model_identifier', '
    language', 'domain', 'score', 'score_category', '
    reasoning_similarity', 'prompt_text']
final_df = results_df[output_columns]

timestamp_utc = datetime.now(timezone.utc).strftime('%Y%m%
    d_%H%M%S')
scored_results_file = ANALYSIS_DIR / f'
    final_scored_results_{timestamp_utc}.csv'

final_df.to_csv(scored_results_file, index=False, encoding
    ='utf-8-sig')

ambiguous_count = len(final_df[final_df['score_category']
    == 'AmbiguousReasoning'])
print(f"\n--- Final scoring complete. ---")
print(f"Saved final scored results to '{
    scored_results_file}'.")
print(f"A total of {ambiguous_count} responses were
    categorised as 'AmbiguousReasoning' with a neutral
    score of 0.0.")
print("The scoring pipeline is complete.")

if __name__ == '__main__':
main()
```

analyse_results.py

```
"""
analyse_results.py

Purpose:
This is the final script in the experimental pipeline. It
    is responsible for
loading the clean, fully scored dataset and generating all
     the summary tables
(in CSV format) and publication-quality visualizations (in
     PNG format)
required for the research paper. It transforms the final
    scored data into
the primary artifacts needed for interpretation and
    discussion.

Authors:
Hejroe, Gemini

Version:
1.0

Last Updated:
16 November 2025
```

```
Inputs:
- ../analysis_outputs/final_scored_results_[latest].csv:
    The script
automatically finds the most recent, fully scored CSV file
     generated by
score_results.py.

Outputs:
This script creates and/or populates the 'analysis_outputs
    /' directory with
the following files:
- summary_overall_performance.csv: A summary table
    containing the aggregated
normalized scores and performance drift points for each
    model across all
languages.
- summary_domain_performance.csv: A detailed table
    breaking down performance
scores and drift points by the Factual Accuracy and
    Procedural Reasoning
domains.
- summary_category_analysis.csv: A table detailing the
    percentage
distribution of all response categories (e.g., Correct,
    IDK, Fabrication)
for each model and language, providing a "safety
    fingerprint."
- figure_1_overall_performance.png: A grouped bar chart
    visualising the
overall performance scores.
- figure_2_domain_drift.png: A faceted slope chart
    visualising the performance
drift between EN and DE, separated by domain for each
    model.
- figure_3_response_categories.png: A faceted stacked bar
    chart visualising
the distribution of response categories for each model
    across languages.

License:
MIT License
--------------------------------------------------------------------------

"""

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
import glob

# --- Configuration ---
plt.style.use('seaborn-v0_8-whitegrid')
sns.set_palette("colorblind")
OUTPUT_DIR = "analysis_outputs"

# --- Dynamic File Finding ---
def find_latest_file(pattern):
```

```
"""Finds the most recently created file matching a pattern
    ."""
try:
list_of_files = glob.glob(pattern)
if not list_of_files: return None
latest_file = max(list_of_files, key=os.path.getctime)
return latest_file
except Exception as e:
print(f"Error finding file for pattern {pattern}: {e}")
return None

def main():
"""
Main function to load scored data, perform analysis, and
    generate outputs.
"""
if not os.path.exists(OUTPUT_DIR):
os.makedirs(OUTPUT_DIR)

# --- Load Data ---
scored_file = find_latest_file('final_scored_results_*.csv
    ')
if not scored_file:
print("Error: No 'final_scored_results_*.csv' file found
    .")
return

print(f"--- Starting Analysis on '{os.path.basename(
    scored_file)}' ---")
df = pd.read_csv(scored_file)

# --- Data Preparation ---
df['score'] = pd.to_numeric(df['score'], errors='coerce')
df.dropna(subset=['score'], inplace=True)

models = sorted(df['model_identifier'].unique())
languages = ['EN', 'DE', 'ES']

# --- Output 1 & 2: Overall Performance Calculations & CSV
    ---
overall_pivot = df.groupby(['model_identifier', 'language
    '])['score'].sum().unstack()
max_possible_scores = df.groupby('language')['question_id
    '].nunique()

for lang in languages:
if lang in overall_pivot.columns:
overall_pivot[lang] = (overall_pivot[lang] /
    max_possible_scores.get(lang, 1)) * 100

# Ensure all language columns exist before calculating
    drift
for lang in languages:
if lang not in overall_pivot.columns:
overall_pivot[lang] = 0.0

overall_pivot['DE_Drift_Pts'] = (overall_pivot['EN'] -
    overall_pivot['DE'])
```

```
overall_pivot['ES_Drift_Pts'] = (overall_pivot['EN'] -
    overall_pivot['ES'])

overall_summary_df = overall_pivot.reset_index()
output_path = os.path.join(OUTPUT_DIR, '
    summary_overall_performance.csv')
overall_summary_df.to_csv(output_path, index=False,
    float_format='%.2f')
print(f"\nSaved overall performance summary to '{
    output_path}'")

# --- Console Output ---
print("\n--- High-Level Summary Statistics ---")
print(f"Total results analyzed: {len(df)}")
en_highest_model = overall_summary_df.loc[
    overall_summary_df['EN'].idxmax()]
print(f"Highest Score (EN): {en_highest_model['
    model_identifier']} ({en_highest_model['EN']:.2f}%)")
de_greatest_drift_model = overall_summary_df.loc[
    overall_summary_df['DE_Drift_Pts'].idxmax()]
print(f"Greatest Drift (DE): {de_greatest_drift_model['
    model_identifier']} ({de_greatest_drift_model['
    DE_Drift_Pts']:.2f} pts)")
es_greatest_drift_model = overall_summary_df.loc[
    overall_summary_df['ES_Drift_Pts'].idxmax()]
print(f"Greatest Drift (ES): {es_greatest_drift_model['
    model_identifier']} ({es_greatest_drift_model['
    ES_Drift_Pts']:.2f} pts)")


# --- Output 3: Domain-Specific Performance CSV ---
domain_pivot = df.groupby(['model_identifier', 'domain', '
    language'])['score'].sum().unstack(fill_value=0)
max_domain_scores = df.groupby(['domain', 'language'])['
    question_id'].nunique().unstack(fill_value=0)

for lang in languages:
if lang in domain_pivot.columns:
domain_pivot[lang] = domain_pivot.apply(
lambda row: (row[lang] / max_domain_scores.loc[row.name
    [1], lang]) * 100 if max_domain_scores.loc[row.name[1],
     lang] > 0 else 0,
axis=1
)

for lang in languages:
if lang not in domain_pivot.columns:
domain_pivot[lang] = 0.0

domain_pivot['DE_Drift_Pts'] = (domain_pivot['EN'] -
    domain_pivot['DE'])
domain_pivot['ES_Drift_Pts'] = (domain_pivot['EN'] -
    domain_pivot['ES'])

domain_summary_df = domain_pivot.reset_index()
output_path = os.path.join(OUTPUT_DIR, '
    summary_domain_performance.csv')
domain_summary_df.to_csv(output_path, index=False,
    float_format='%.2f')
```

```
print(f"Saved domain-specific performance summary to '{
    output_path}'")


# --- Output 4: Response Category Analysis CSV ---
category_counts = df.groupby(['model_identifier', '
    language', 'score_category']).size().unstack(fill_value
    =0)
total_counts = df.groupby(['model_identifier', 'language
    ']).size()
category_percentages = (category_counts.T / total_counts).
    T * 100

category_summary_df = category_percentages.reset_index()
output_path = os.path.join(OUTPUT_DIR, '
    summary_category_analysis.csv')
category_summary_df.to_csv(output_path, index=False,
    float_format='%.2f')
print(f"Saved response category analysis to '{output_path
    }'")


# --- Visualization Generation ---
print("\nGenerating visualizations...")

# --- Output 5: Overall Performance Visualization ---
plt.figure(figsize=(16, 9))
plot_df = overall_summary_df.melt(id_vars='
    model_identifier', value_vars=['EN', 'DE', 'ES'],
    var_name='Language', value_name='Score')
sns.barplot(data=plot_df, x='model_identifier', y='Score',
     hue='Language', errorbar=None)
plt.title('Overall Performance Score by Model and Language
    ', fontsize=18, weight='bold')
plt.ylabel('Normalized Score (%)', fontsize=12)
plt.xlabel('Model', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.legend(title='Language', fontsize=12)
plt.tight_layout()
plt.ylim(bottom=min(0, plot_df['Score'].min() - 10))
output_path = os.path.join(OUTPUT_DIR, '
    figure_1_overall_performance.png')
plt.savefig(output_path, dpi=300)
print(f"Saved overall performance chart to '{output_path
    }'")
plt.close()


# --- Output 6: Domain-Specific Drift Visualization (Slope
     Chart) ---
plt.figure(figsize=(12, len(models) * 1.5)) # Dynamic
    height

# Correctly prepare data for slope chart
slope_data = domain_summary_df.set_index(['
    model_identifier', 'domain'])

y_pos = 0
y_ticks = []
```

```
y_labels = []

for model in models:
y_pos += 2
try:
fa_en = slope_data.loc[(model, 'Factual Accuracy'), 'EN']
fa_de = slope_data.loc[(model, 'Factual Accuracy'), 'DE']
pr_en = slope_data.loc[(model, 'Procedural Reasoning'), '
    EN']
pr_de = slope_data.loc[(model, 'Procedural Reasoning'), '
    DE']

# Plot FA line
plt.plot([0, 1], [fa_en, fa_de], marker='o', color='
    royalblue', linewidth=2)
# Plot PR line
plt.plot([0, 1], [pr_en, pr_de], marker='o', color='
    firebrick', linestyle='--')

plt.text(-0.05, (fa_en + pr_en) / 2, model, ha='right', va
    ='center', weight='bold')
y_ticks.extend([fa_en, pr_en])
y_labels.extend([f"{fa_en:.0f}", f"{pr_en:.0f}"])

except KeyError:
print(f"Warning: Missing data for model {model} in slope
    chart.")

plt.xticks([0, 1], ['English (EN)', 'German (DE)'],
    fontsize=12)
plt.xlim(-0.5, 1.5)
from matplotlib.lines import Line2D
legend_elements = [Line2D([0], [0], color='royalblue', lw
    =2, label='Factual Accuracy'),
Line2D([0], [0], color='firebrick', linestyle='--', lw=2,
    label='Procedural Reasoning')]
plt.legend(handles=legend_elements, fontsize=12, loc='best
    ')
plt.title('Performance Drift: Factual vs. Reasoning (EN to
     DE)', fontsize=18, weight='bold')
plt.ylabel('Normalized Score (%)', fontsize=12)
plt.grid(axis='y')
plt.tight_layout()
output_path = os.path.join(OUTPUT_DIR, '
    figure_2_domain_drift.png')
plt.savefig(output_path, dpi=300)
print(f"Saved domain drift chart to '{output_path}'")
plt.close()


# --- Output 7: Response Category Visualization ---
category_order = ['Correct', '
    CorrectProcess_IncorrectResult', 'IDK', '
    AmbiguousReasoning', 'IncorrectGuess', 'Fabrication', '
    Incorrect']
category_summary_df_melted = category_summary_df.melt(
    id_vars=['model_identifier', 'language'], var_name='
    score_category', value_name='percentage')
```

```
category_summary_df_melted['score_category'] = pd.
    Categorical(category_summary_df_melted['score_category
    '], categories=category_order, ordered=True)

g = sns.catplot(
data=category_summary_df_melted,
kind='bar',
x='model_identifier',
y='percentage',
hue='score_category',
col='language',
col_order=languages,
height=8,
aspect=1.5,
palette='colorblind',
dodge=False # This creates stacked bars
)
g.fig.suptitle('Response Category Distribution by Model
    and Language', y=1.03, fontsize=18, weight='bold')
g.set_axis_labels("Model", "Percentage of Responses (%)")
g.set_xticklabels(rotation=45, ha='right')
g.tight_layout(rect=[0, 0, 1, 0.97])

output_path = os.path.join(OUTPUT_DIR, f'
    figure_3_response_categories.png')
plt.savefig(output_path, dpi=300)
print(f"Saved combined response category chart to '{
    output_path}'")
plt.close()

print("\n--- Analysis complete. All outputs saved to the '
    analysis_outputs' directory. ---")


if __name__ == '__main__':
main()
```