Evaluating the Meta- and Object-Level Reasoning of Large Language Models for Question Answering

Anonymous submission

Abstract

Large Language Models (LLMs) excel in natural language tasks but still face challenges in Question Answering (QA) tasks requiring complex, multi-step reasoning. We outline the types of reasoning required in some of these tasks, and reframe them in terms of *meta-level reasoning* (akin to high-level strategic reasoning or planning) and *object-level reasoning* (embodied in lower-level tasks such as mathematical reasoning). FRANKLIN, a novel dataset with requirements of meta- and object-level reasoning, is introduced and used along with three other datasets to evaluate four LLMs at question answering tasks requiring multiple steps of reasoning. Results from human annotation studies suggest LLMs demonstrate meta-level reasoning with high frequency, but struggle with object-level reasoning tasks in some of the datasets used. Additionally, evidence suggests that LLMs find the object-level reasoning required for the questions in the FRANKLIN dataset challenging, yet they do exhibit strong performance with respect to the meta-level reasoning requirements.

1 Introduction

Large Language Models (LLMs) have emerged as general-purpose natural-language-based task solvers. Earlier models such as BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019) demonstrated capability at tasks previously performed by task-specific models, such as sentiment analysis, while today's vastly increased model sizes and context windows means LLMs now find application in tasks which handle large amounts of data such as question answering (QA) over large volumes of data (Guu et al. 2020), and reading comprehension (Kočiský et al. 2018). Alongside advances in model size, turn-based conversation has become a key mode of interaction, with commercial products like ChatGPT bringing considerable non-expert attention to the field. QA is a primary function of these LLM-based assistants, with research efforts shifting away from simpler factoid questions and towards more complex QA varieties which require reasoning in a human-like manner (e.g., StrategyQA (Geva et al. 2021), CommonsenseQA (Talmor et al. 2019).) However, for a model to perform well at these QA tasks, it is not necessarily enough to simply ask a given question of a model – supplementary techniques such as *Chain-of-Thought* (Wei et al. 2022) are required to elicit more complex reasoning.

In this paper, we discuss the reasoning tasks expected of LLMs (section 2.1), and some of the methods for improving LLM performance on reasoning tasks in section 2.2. We introduce our re-framing of the LLM reasoning discourse in section 3.1 in terms of the ability of LLMs to demonstrate meta- and object-level reasoning, and introduce our novel FRANKLIN dataset in section 3.2. We then introduce and describe two annotation studies in section 4.1, which were conducted to evaluate the reasoning capabilities of a range of state-of-the-art LLMs described in section 4.3. These models were used to generate responses for a range of QA datasets, whose requirements we re-frame in terms of meta- and object-level reasoning in 4.2. Our research questions, shown below, are addressed using results from our annotation studies in section 5.

RQ1 Do LLMs demonstrate object-level reasoning?

RQ2 Do LLMs demonstrate meta-level reasoning?

RQ3 Does our novel FRANKLIN dataset present a challenge for LLMs?

Our contributions are:

- A re-framing of existing LLM reasoning discourse in terms of meta- and object-level reasoning (section 3.1), enabling better classification of their strengths and weaknesses.
- The introduction of the novel FRANKLIN dataset, which contains meta- and object-level reasoning requirements.
- Evaluation of a range of state-of-the-art LLMs on datasets requiring multi-step reasoning, and discussion of the strengths and limitations of LLMs (section 4).
- A claim that LLMs generally do not possess sufficient object-level reasoning to widely succeed at the datasets evaluated.
- An additional claim that LLMs are able to demonstrate meta-level reasoning consistently across the variety of datasets selected.
- Finally, a claim that our FRANKLIN dataset presents a challenge to LLMs through discussion of the low rates with which answers are provided, and the types of errors that LLMs make.

2 Background

We overview a selected range of reasoning tasks which LLMs are evaluated against, and techniques by which increased performance is extracted from LLMs.

2.1 Reasoning Tasks

The term *reasoning* is a broad cognitive concept with many forms. Reasoning encompasses the drawing of a conclusion, using logical laws, from a set of statements (formally (Bundy 1983) or informally (Wason and Johnson-Laird 1972)); incorporates elements of attitude revision (McHugh and Way 2018); and may be intuitive or explicit (Sloman 1996). Our working definition of reasoning, aiming to take into account the broad range of tasks to which the term is applied, is *a task which requires some operation to infer conclusions from a set of premises*. The range of tasks on which LLMs are evaluated represents the breadth of application of the term, with key tasks including common sense reasoning, mathematical reasoning, and symbolic reasoning. We are particularly interested in *multi-step* reasoning, which requires multiple intermediate steps of inference to draw a final conclusion.

Common sense reasoning concerns knowledge about everyday concepts which is *generally accepted* by a majority of people (Bhargava and Ng 2022). While the notion of a formal common sense logic does exist (Booth, Meyer, and Varzinczak 2012), we will discuss *informal* common sense reasoning grounded in natural language. LLMs have been shown to reflect human beliefs about generic concepts across a range of domains (Weir, Poliak, and Durme 2020), and reason about physical properties of everyday objects and situations (Bisk et al. 2020; Goel, Feng, and Boyd-Graber 2019). Similarly, LLMs encode relational data, allowing recalling of facts in a similar manner to symbolic knowledge bases (KBs) (Petroni et al. 2019). These instances reflect simple tasks where the knowledge, implicit in the parameters of an LLM, can be recalled. However, common sense reasoning can also be a requirement for some multi-step reasoning tasks, such as the creation of a strategy for answering a question which requires multiple inferences (Geva et al. 2021). As we will see, natural language-based common sense reasoning is a requirement for a variety of tasks which require multiple steps of inference to achieve a wider goal.

Mathematical reasoning concerns a model's ability to perform mathematical operations to solve problems (Ahn et al. 2024). Specific tasks include arithmetic reasoning, such as addition and division, which can be expressed simply in symbolic form (Yuan et al. 2023), or in longer-form, text-based problems (Cobbe et al. 2021; Hendrycks et al. 2021). Geometry problems, which represent a conceptually harder challenge, are another example of problems requiring mathematical reasoning (Chen et al. 2021).

Symbolic reasoning tasks involve performing an action according to formal rules, albeit imitated using the prompting and output of an LLM. This is a broader task than the mathematical reasoning task, which encompasses arithmetic and polynomial evaluation. Wei et al. (2022) describe two tasks which illustrate the challenge, although it is noted that these toy tasks are within current abilities of LLMs. In *last letter concatenation*, a model concatenates the last letters of a full name (e.g., *Barack Obama* $\rightarrow$ *ka*), and in *coin flipping*, models are prompted to output the state of the coin after given an initial state and a number of flips. Other examples include the emulation of formal deductive reasoning in natural language (Han et al. 2024; Clark, Tafjord, and Richardson 2020). Some symbolic reasoning tasks are presented to the model not expressed in natural language, but symbolically. For example, finding checkmate in a chess game – one of the many symbolic reasoning tasks in BIG-bench (BIG-bench 2023).

There is debate over whether LLMs are *actually* reasoning rather than emulating or imitating it, which itself is part of a wider debate of whether LLMs truly understand language and meaning. Even when LLMs appear to perform reasoning tasks, it is not clear that they are reliant upon reasoning (Wei et al. 2022), or if they simply using heuristics to make predictions (Patel, Bhattamishra, and Goyal 2021).

Many of these reasoning tasks are exemplified in QA datasets, against which LLMs are evaluated. They also do not exist in isolation, and often require multiple steps of inference. For example, multi-step common sense reasoning is employed to infer the steps required for solving the natural language problems in the GSM8k (Cobbe et al. 2021) and MATH (Hendrycks et al. 2021) datasets. Similarly, basic mathematical reasoning about, for example whether a given date is before or after another date, is required for many questions in the StrategyQA dataset (Geva et al. 2021).

2.2 Improving LLM Performance at Reasoning Tasks

Techniques for achieving improved performance with LLMs on reasoning tasks can broadly be placed into two categories: fine-tuning and prompt-based approaches.

Fine-tuning is a paradigm which involves taking a pre-trained model and updating the model's parameters by further training on a specific dataset (Devlin et al. 2019). Fine-tuning is performative in increasing the performance of LLMs in a variety of reasoning techniques, such as arithmetic reasoning Cobbe et al. (2021); Hendrycks et al. (2021) and common sense QA tasks (Talmor et al. 2019).

Prompting-based techniques, sometimes referred to as *in-context learning*[1] involve taking a model's frozen weights and manipulating the content of the prompt to 'externalise' the model's reasoning in the form of natural language. *Chain-of-Thought* (CoT, Wei et al. (2022)) involves inserting intermediate natural language reasoning steps into the process of solving common sense, mathematical, and symbolic reasoning tasks. A variety of closely-related techniques have since appeared and shown to increase the performance of models on reasoning tasks. Such techniques generally involve forcing a model to generate the intermediate reasoning steps required to compute the solution to a multi-step reasoning problem. Examples include appending *"Let's think step-by-step"* to the end of a question (Kojima et al. 2022); instructing models to decompose the problem into

[1] Erroneously, as no gradient update/learning takes place.

sub-problems (Zhou et al. 2023); and allowing models to explore multiple reasoning paths (Wang et al. 2023).

Aside from these *Chain-of-Thought*-esque prompt formats, models are also prompted in an *n-shot* setting in bring about improved performance. This involves providing the model with a number of demonstrations of the task at inference time, that is, in the input to the model itself (Brown et al. 2020). For example, if instructing a model to provide English to German translations, one might include n example English-German sentence pairs in the prompt itself. n-shot prompting has been shown to improve the performance of models at variety of tasks (Radford et al. 2019; Brown et al. 2020) and indeed forms a standard component of the evaluation of LLMs on certain datasets. For example, the performance of Llama 3.1 on GSM8k is described in an 8-shot setting, with some form of Chain-of-Though prompting applied.

3 Reframing the Reasoning Task

We now describe *meta- and object-level reasoning* and their relevance to the task at hand, namely, multi-step QA with LLMs. Then, building on these distinct reasoning types, we introduce a novel dataset, FRANKLIN, which is inspired by the FRANK system, a QA system which employs meta- and object-level reasoning to infer answers to queries.

3.1 Meta- and Object-Level Reasoning

Meta- and object-level reasoning are terms associated with symbolic AI, particularly the automated reasoning and proof planning domains. We will first describe a range of definitions of these two concepts to build up a picture of their meaning.

Formally, in automated reasoning, *meta-level* reasoning refers to the reasoning about the representation of a theory, while the theory itself is at the *object-level* (Bundy 1983). Bundy, Byrd, and Luger (1979) uses meta-level inference to control the search of a solution to mechanics problems phrased in natural language, while object-level inference is used to compute the steps of the solution itself. Christodoulou and Keravnou (1998) describes the role of meta-level reasoning as planning problem-solving strategies, controlling the use of different *problem solvers* (which can be thought of as object-level reasoning components), and notes the use of meta-level reasoning in adapting a strategy to new knowledge which may arise during computation. Aiello, Nardi, and Schaerf (1991) describe meta-level reasoning as *reasoning about reasoning*, and also note its functionality in driving search strategies and the modification of a system's own behaviour. They also, in the context of an agent-based system, distinguish the meta- and object-levels by stating that agents' world knowledge is on the object-level, while meta-level knowledge governs the links between different agents. Genesereth (1983) distinguishes the actions of an AI system as *base-level* (or, object-level) and meta-level. Object-level actions achieve the program's goals, while meta-level actions decide which object-level actions to perform. Nuamah and Bundy (2023) introduce a formalism for representing knowledge in a QA system consisting of attribute-value pairs. This formalism introduces additional attributes to the standard $\langle$*subject*,*predicate*,*object*$\rangle$ triple, which may be meta- or object-level attributes. Object-level attributes are those which encode the meaning of a factual statement, such as *subject* and *predicate*, while meta-level attributes capture meta-information, such as the data source for a given fact.

To summarise the above examples, meta-level reasoning approximately corresponds to the high-level planning of a solution to a problem, the decomposition of a problem into intermediate steps, and the decisions on which sub-components of a system to employ to achieve a specific task. Reasoning on the object-level concerns the application of the sub-components. This includes lower-level inferences, such as mathematical operations or natural language deductions, which are required to execute the intermediate steps.

We find that this delineation of reasoning tasks provides meaningful detail and structure to the discourse and classification of the reasoning tasks embodied in multi-step QA datasets on which LLMs are evaluated. Taking GSM8k as an example, it is described as embodying a single reasoning task, namely *mathematical reasoning*. However, cursory analysis of the problems contained within the dataset show that both meta- and object-level reasoning are required to correctly compute answers to the questions. In section 4.2 we describe further examples of datasets which require meta- and object-level reasoning, showing that our categorisation of reasoning types as meta- or object-level generalises to a range of QA tasks, in addition to adding more fine-grained meaning. Questions in these QA datasets are the basis for our annotation studies and evaluation of the meta- and object-level reasoning of the range of LLMs selected in 4.3. However, in conducting our studies on the ability of LLMs to demonstrate meta- and object-level reasoning, we do not claim here that LLMs have any formal meta- or object-level reasoning component, and stress here that when applying these terms to the evaluation of LLMs, we are not evaluating a formal meta- or object-level reasoning component. Rather, when we refer to LLMs as *demonstrating* meta- or object-level reasoning, we refer to their ability to *emulate*, or *imitate* such processes via their text generation paradigm. Our interpretation of the terms meta- and object-level reasoning is summarised below.

Meta-level reasoning *High-level planning*. With LLMs, this is demonstrated and embodied in an informal, natural language-based decomposition of a problem in to sub-problems or intermediate steps.

Object-level reasoning *Low-level execution*. With LLMs, this is demonstrated in the execution of intermediate steps created by the meta-level reasoning process. Execution of these steps may require a specific task, for example, mathematical reasoning.

It is from this characterisation which two of our research questions, introduced in section 1, are drawn. We revisit them here and give further detail using our above definitions.

RQ1 *Do LLMs demonstrate object-level reasoning?* Object-level reasoning involves low-level inferences, such as the execution of mathematical operations or

natural language deduction using common sense knowledge. Can LLMs demonstrate a ability at this task across a range of datasets?

RQ2 *Do LLMs demonstrate meta-level reasoning?* Meta-level reasoning governs the high-level planning and strategy for finding a solution to a problem. While we do not pretend that LLMs are employing some formal meta-level process, can LLMs demonstrate an ability to plan a solution to a range of problems as embodies in the range of datasets selected?

3.2 Introducing FRANKLIN

The FRANK System To give further example of meta- and object-level reasoning processes in a QA setting, we will refer to the FRANK system (Nuamah and Bundy 2020) as an example. FRANK is a QA system in the form of a symbolic reasoning framework which employs meta- and object-level reasoning in the form of a set of rules (Bundy and Nuamah 2022). These rules break queries down into sub-problems; collect data from online knowledge sources such as Wikidata, and apply mathematical operations over that data. In FRANK, meta-level reasoning governs the high-level approach to answering a question, including the deduction of which intermediate inferences and operations are necessary. Object-level reasoning manifests in both the queries to knowledge bases, and in the mathematical operations applied to the data returned from knowledge bases (with the decision to use such operations taking place at the meta-level.) Multiple lines of reasoning may be explored by the system before a final solution is assembled – reasoning is dynamic at inference time, and not pre-determined for a given question type.

FRANK's functionality illustrated using questions concerning the values of geopolitical indicators belonging to different countries and regions at various points in time. As an example, consider the question: "*Which country in Africa had the lowest population in 2012?*". This cannot necessarily be answered as a factoid style question by retrieving a value from a knowledge base because multiple steps of inference are required, in contrast to a question like "*What is the capital of Ghana?*", which is simply a single fact that can be looked up. One solution FRANK may explore is to split *Africa* into constituent countries, search for their populations in 2012, and compare values to find an answer.

The symbolic nature of this system does lead to limitations, generally as a result of the levels of hand-engineering required. Although solutions are not hard-coded, rules which decompose queries, perform information retrieval, and aggregate data do require hand-engineering. This lends part of the overall motivation to the project: exploring the capability of LLMs at functionality that can be performed by explicit symbolic components.

The FRANKLIN Dataset Given this task of meta- and object-level reasoning, we introduce a novel dataset inspired by the FRANK system and its exemplar domain of geopolitical indicators. This dataset, which we call FRANKLIN (FRANK Library of Ideal Narratives) [2], consists of questions, paired with template-based, natural language, step-by-step descriptions modelled on how FRANK would nominally decompose a problem using formal deductive reasoning. Four question templates make up the dataset, shown in figure 1.

A. Future prediction What will be the `<property>` of `<subject>` in `<future_year>`?

B. Region comparison Which country in `<region>` had the `<operator>` `<property>` in `<past_year>`?

C. Past comparison & future prediction In `<future_year>`, what will be the `<property>` of the country in `<region>` which had the `<operator>` `<property>` in `<past_year>`?

D. Future prediction & comparison Will `<subject_A>` or `<subject_B>` have a `<operator>` `<property>` in `<future_year>`?

Figure 1: The four question types which make up the FRANKLIN dataset.

Values which slots may take are detailed in table 1, and the resulting number of possible instantiations are given in table 2. An instantiated example of type B: region comparison is shown in figure 2. Our initial proof-of-concept release contains 400 examples, with 100 examples for each question type.

Which country in Eastern Europe had the highest energy consumption in 2019?

1. A list of countries located in Eastern Europe was needed. ***Meta.***
2. 10 countries were found in Eastern Europe, including Hungary, Romania and Slovakia. ***Object.***
3. The energy consumption for each of these countries in 2019 was needed for a comparison. ***Meta.***
4. Data on each country's energy consumption in 2019 was found. ***Object.***
5. The values of energy consumption were compared to each other. ***Object.***
6. The answer to the question is the country which had the highest value. ***Meta.***

Figure 2: Example of the *region comparison* question type from the FRANKLIN dataset. ***Step reasoning type is indicated in bold and italics.***

The natural language explanations that accompany the questions in the dataset are inspired by the functionality of

[2] Available in a proof-of-concept alpha version at the anonymised Github link https://anonymous.4open.science/r/aaai2024-llm4plan-anon-repo-link/.

Field	Description	Number available	Example(s)
<property>	Geopolitical indicator	8	*Female population*
<subject>	Country (using ISO 3166 standard).	249	*Ghana*, *France*
<region>	ISO 3166 'sub-region'	16	*Western Europe*
<{future,past}_year>	Year in range [2008, 2030].	32	*2009*, *2027*
<operator>	Comparison operation	2	*Maximum*, *minimum*

Table 1: Explanation of slots which can be instantiated in FRANKLIN question templates.

Question type	Possible instances
A. Future pred.	3.19×10^4
B. Region comp.	4.10×10^3
C. Past comp. & future pred.	5.24×10^5
D. Region comp.	1.59×10^7

Table 2: Number of possible instantiations for each question type in FRANKLIN.

FRANK, and spell out the reasoning required of any system tackling the problem, with each step having a label of meta- or object-level reasoning. The reasoning type label for each step was a product of an annotation task completed by the authors. As discussed above, meta-level reasoning is required to plan out how an answer to the question can be found, comprising of the setting of sub-goals and intermediate steps, and how numeric data may be aggregated to estimate an answer. In parallel, object-level reasoning is required to retrieve the data, and perform mathematical operations. The information retrieval aspect requires recalling accurate numeric data, while mathematical operations required range from simple comparisons to multiplication with 5-7 digit numbers. For these reasons, expect both applications of object-level reasoning to be challenging for LLMs.

It should be noted that, in its current early version, the step-by-step content paired with each question takes the form of an explanation phrased as if communicating a process *after* it has been performed, rather than planning out a process to be performed. Additionally, the steps which indicate that object-level reasoning has taken place do not spell out the actual operations performed, but only allude to the fact that they have been performing. These limitations, and others, will be the subject of future work as discussed in section 6.

The FRANKLIN dataset forms the basis for our third research question, which we repeat and expand on here.

RQ3 *Does our novel* FRANKLIN *dataset present a challenge for LLMs?* Meta-level reasoning is required to plan out a solution to the question in terms of the necessary intermediate inferences, while object-level reasoning requires recalling factual information to high precision, and mathematical operation on said data. Can LLMs perform these actions to a sufficiently high degree?

4 Experiment Design

In this section, we describe the annotation studies which were conducted. We describe the content of the studies themselves; list the datasets and models which were used to generate materials for the study; and finally our evaluation metrics which we use to evaluate our research questions in section 5.

4.1 Annotation Studies

We firstly describe the design of the studies themselves, including the questions which were presented to participants, and the size of the studies in terms of the number of examples annotated.

Study Design Two online human annotation studies in which we evaluated the ability of a range of LLMs to demonstrate meta- and object-level reasoning. Four datasets, described in 4.2 were selected. Responses to a random sample of the questions in these datasets were generated using four LLMs, introduced in section 4.3. For study 1, we prompted models to generate *answers* to a given question, with the intention of observing object-level ability. For study 2, we first prompted models to generate *plans* for finding an answer to a given question, and followed up with an instruction to execute the plan step-by-step. This study was designed to observe the models' meta-level reasoning ability, and also the influence of the breaking down of a problem on their ability to produce answers. Full details of the prompts used are given in appendix A. In both annotation studies (built with Qualtrics[3]), we asked human participants sourced from Prolific[4] to answer questions about models' responses on a 5-point Likert scale: *strongly disagree*, *somewhat disagree*, *neither agree nor disagree*, *somewhat agree*, and *strongly agree*. In study 1, for each question, participants are required to respond to the statements below.

1. The response contains an answer to the question.
2. The response contains a clear step-by-step plan.
3. I would be satisfied with the response if I had asked the question.

Similarly, the below list shows the statements presented for each example in study 2.

1. The response takes a rational approach to answering the question.
2. The response contains an answer to the question.

[3]https://www.qualtrics.com/

[4]https://www.prolific.com/

3. The response contains a clear step-by-step plan of how an answer can be found.
4. Each step in the plan is visibly performed.

Study Size 64 examples were randomly selected from the test split of each of the four datasets, and responses were generated for each example with each of the four models. For structured responses like step-by-step plans, models often generate Markdown formatting such as bold text and bulleted lists. We converted this formatting to HTML, which is supported by Qualtrics, so that such formatting was visible to users. We also cleaned responses of LaTeX maths formatting to aid legibility. This gives a total of $4\times4\times64$=1,024 examples. 256 participants were recruited for each study, with a pre-screening process requiring participants' first language to be English. Each participant annotated 16 examples – one for each model/dataset combination. This resulted in 256×16 annotations evenly distributed over our 16 model/-dataset combinations giving 4 annotations per example. A pilot study was conducted with 32 participants in which 8 annotations per example were collected, however analysis of participant agreement in terms of the Standard Error of the Mean (SEM) showed that 4 annotations per example were sufficient.

4.2 Dataset selection

We selected datasets embodying tasks which require both meta- and object-level reasoning to perform.

GSM8k (Cobbe et al. 2021) contains grade school mathematics problems formulated in natural language, requiring meta-level reasoning to plan the step-by-step approach, and object-level reasoning to perform the arithmetic itself. **StrategyQA** (Geva et al. 2021) contains questions in which the inference requirements are said to be implicit in the question. Meta-level reasoning is required to decompose the problem and decide which intermediate inferences are necessary, while object-level reasoning is required to make deductions about relevant facts. **HotpotQA** (Yang et al. 2018) requires meta-level reasoning to plan the intermediate steps in answering the question, and object-level reasoning to synthesise facts into an answer. We also include the novel **FRANKLIN** dataset, introduced above, which requires meta-level reasoning to decompose problems into sub-problems, and object-level reasoning to retrieve information and perform mathematical reasoning. An example of the questions in each dataset is given in figure 3.

To the best of our knowledge, according to the models' white papers (Abdin et al. 2024; Dubey et al. 2024; Team et al. 2024), none of these datasets were part of a given model's pre-training or fine-tuning data, and therefore, not used to train the models themselves.

4.3 Model selection

Four off-the-shelf, pre-trained models were used without fine-tuning. Meta's Llama 3.1 8B (Dubey et al. 2024), Microsoft's Phi 3.5 Mini (Abdin et al. 2024), and Google's Gemma 2 9B (Team et al. 2024) were selected as examples of popular, performative, open-source models targeted towards QA and reasoning. OpenAI's GPT-4o-mini, a closed-source model and smaller version of the flagship GPT-4o, was used as an additional comparison. The Meta, Microsoft and Gemma models were downloaded from Huggingface and run on a local compute cluster, while the OpenAI model was queried through the OpenAI API. Models were run in their default configurations aside from a reduction in temperature. Temperature is a generation parameter which takes a positive number, typically on the order of 10^0. It is proxy for 'creativeness', with higher temperatures of ¿1 leading to 'more creative and inspiring' outputs (Dubey et al. 2024). We use a temperature of 0.2 to generate more rational, less creative responses.

FRANKLIN For the country in Western Europe that had the lowest GDP in 2011, what will be its urban population in 2027?

GSM8k Peter has twice as many socks as Jack and half times as many dishes as jack. Jack collected twice as many dishes as socks in the store. If jack collected 60 dishes, calculate the total number of socks and dishes they have together

HotpotQA Jaroslav Navratil plays for a football club that was founded in what year?

StrategyQA Would it be common to find a penguin in Miami?

Figure 3: Example questions from each of the datasets employed in the study.

4.4 Metrics

In assessing our claims, we refer to the metrics outlined here.

Answer Failure Rate (AFR) In studies 1 and 2, we asked participants to indicate whether a given response contained an answer to the question at hand (shown in questions 1 and 2 for studies 1 and 2 respectively). The AFR is derived from these results. It shows, for a given model/dataset combination, the proportion of questions which contain **no attempted answer** to the question at hand. We focus on AFR rather than a standard accuracy metric because we want to observe an upper bound for a model's object-level reasoning ability – we are less interested in absolute performance on a dataset, more so in making a comment about whether sufficient object-level reasoning is demonstrated.

To arrive at the AFR for the responses for each model/-dataset combination, we took the following approach. For each set of four annotations for a given response, we mapped the ratings on the 5-point Likert scale to a 3-point scale representing *strongly/somewhat disagree*, *neither agree nor disagree*, and *strongly/somewhat agree*. We then took a majority vote of these 3-point ratings to achieve a single verdict for given response. The proportion of non-*strongly/somewhat agree* verdicts gives the AFR.

Rational Approach Rate (RAR) In study 2, the focus is on the ability of LLMs to generate plans and, if possible,

execute them to produce an answer. In question 1 of study 2, we asked participants to indicate whether a rational approach to answering the question was present. RAR denotes the proportion of questions which, according to annotators, contained a rational approach to solving the problem in the response. RAR rewards models for demonstrating a general understanding of the steps required to solve a problem, even if a step-by-step plan is not explicitly created. Given that our response generation procedure was unconstrained, we did not want to excessively penalise models for not adhering to the 'step-by-step' instruction.

As with the process for obtaining AFR above, we map our results to a 3-point Likert scale and take a majority vote. This process yields a verdict on whether a response contains a rational approach to answering the question in the case of RAR, and whether a response contains a step-by-step plan in the case of PCR.

Plan Creation Rate (PCR) With PCR, we are again in the setting of study 2 where LLMs were instructed to produce a step-by-step plan as part of their response. In question 3 of study 2, we asked participants to indicate whether a step-by-step plan was present in the response. Similarly to RAR, PCR denotes the proportion of questions which contained a clear step-by step plan in the response. PCR specifically targets the formatting of a step-by-step plan – a stricter requirement than simply producing a rational approach – models were required to clearly format a step-by-step plan.

As with the processes for the above metrics, we mapped 5-point Likert scale response to a 3-point scale, before taking a majority vote to arrive at a single verdict for the presence of a plan in a given response. PCR indicates, the proportion of questions which contained a clear step-by-step plan according to this process.

5 Results and Discussion

We now bring together our findings from our annotation studies, addressing our research questions using the metrics defined in section 4.4.

5.1 Do LLMs demonstrate object-level reasoning?

Table 3 shows AFR for studies 1 and 2. The left number denotes AFR from the study 1, where LLMs provided only an answer to the question. The right figure (shaded) denotes AFR from study 2, where LLMs were instructed to create a plan before executing that plan to answer the question.

The figures show that answers were frequently not present for a variety of model/dataset combinations. In the study 1 setting, models overall found FRANKLIN the harder of the datasets, with GPT 4o-mini performing best with an AFR of 53%. Gemma 2's AFR of 88% on the FRANKLIN dataset was the worst of any model/dataset combination. GSM8k, with its simple arithmetic and verbose question formats, was comparatively easy compared to other datasets, with models (except in the case of Phi 3.5) failing to provide answers for less than 10% of responses. HotpotQA and StrategyQA, with their text-based common sense knowledge requirements, occupied a middle-ground in terms of difficulty.

AFR is consistently lower in the setting of study 2, indicating that the generation of a plan enabled models attempt answers more frequently. This result generally aligns with the results of Chain-of-Thought-adjacent work, in which models are found to achieve better performance when prompted to decompose problems. The exception to this decreased AFR was GPT 4o-mini, which did create plans for the questions, when asked, but specifically declined to execute that plan. We hypothesise that this is the result of safety 'guardrails' being put in place by OpenAI.

We conclude by claiming that there is preliminary evidence that, while instructing the model to perform meta-level reasoning before answering the question results in lower AFR, models did not sufficiently, or consistently, demonstrate high levels of object-level reasoning across the range of multi-step question answering datasets.

5.2 Do LLMs demonstrate meta-level reasoning?

To answer this question, we make use of our RAR and PCR metrics described in 4.4. Table 4 shows RAR and PCR across model/dataset combinations.

Results at this meta-level reasoning task show both stronger, and more consistent levels of performance at the meta-level reasoning task, frequently over 95% for many model/dataset combinations, even for the FRANKLIN dataset. As described in section 4.4 reported RAR to provide an indication that, even if the model does not produce a step-by-step plan, it still approaches the problem in a rational manner according to our annotators. This way, we have evidence that models possess sufficient meta-level reasoning to approach the problem in an interpretable, human-understandable manner – and we can explore ways of imposing greater structure on this in future work.

In contrast to the results for AFR in table 3, results for RAR and PCR appear to suggest that models are very competent in generating solutions to problems which take an approach which humans rate as rational, and they are similarly capable of structuring this approach in a step by step manner. In many cases, models were able to do this for all examples in a dataset, such as in the case of Phi 3.5 on the GSM8k dataset. Although above we suggested that the object-level reasoning in GSM8k was easier for models due its simple arithmetic and verbose questions, and that FRANKLIN was a harder task, we see similar levels of very high competence at the meta-level reasoning task across the range of datasets.

Again, we point out that there is speculation about whether LLMs are *actually* reasoning, rather than simply *imitating* it (Wei et al. 2022). We share this scepticism of models' abilities to formally reason at the meta-level and do not claim that models possess any kind of implicit, underlying, symbolic representation which this process is being completed by. **However, we believe that our results suggest that models are able to *imitate* meta-level reasoning in their text generation paradigm.**

5.3 Does the FRANKLIN dataset present a challenge for LLMs?

Table 3 shows that LLMs clearly struggled with questions from FRANKLIN without first being prompted for a plan,

Dataset	FRANKLIN		GSM8k		HotpotQA		StrategyQA	
Model	S1	S2	S1	S2	S1	S2	S1	S2
google/gemma-2-9b-it	0.88	0.33	0.05	0.19	0.28	0.20	0.25	0.09
meta-llama/Meta-Llama-3.1-8B-Instruct	0.80	0.05	0.08	0.05	0.69	0.19	0.33	0.12
microsoft/Phi-3.5-mini-instruct	0.75	0.16	0.31	0.02	0.52	0.09	0.23	0.05
openai/gpt-4o-mini	0.53	1.00	0.02	0.02	0.05	0.66	0.11	0.72

Table 3: Answer Failure Rate for study 1 (S1) and study 2 (S2, in the shaded boxes). Lower is better. As described in section 4.4, this figure is the answer to the question "*What proportion of responses contained no attempt at an answer to a given question?*"

Dataset	FRANKLIN		GSM8k		HotpotQA		StrategyQA	
Model	RAR	PCR	RAR	PCR	RAR	PCR	RAR	PCR
google/gemma-2-9b-it	0.95	0.88	0.88	0.91	0.84	0.84	0.95	0.94
meta-llama/Meta-Llama-3.1-8B-Instruct	0.95	0.95	0.95	1.00	0.92	0.91	0.95	0.95
microsoft/Phi-3.5-mini-instruct	0.95	0.94	1.00	0.97	0.97	0.92	0.98	0.95
openai/gpt-4o-mini	0.83	0.88	1.00	0.98	0.81	0.86	0.81	0.80

Table 4: Rational Approach Rate from study 2 (Plan Creation Rate in shaded box). Higher is better. PCR answers the question "*What proportion of responses contained a clear step-by-step plan?*", while RAR answers the question "*What proportion of responses outlined a rational approach to answering the question?*".

more so than for other datasets. While the instruction to produce a plan before answering the question lowered AFR, we referred above to analysis of a small sample of answers which show that this lower figure does convince us that models have the necessary object-level reasoning to provide *correct* answers. Different error modes are present when an answer is attempted, including data fabrication, in which the model reported non-existent values which the model claims to have found in knowledge sources; inaccurate or low-precision data, where the model reports heavily rounded or incorrect values; and incorrect arithmetic. Examples are illustrated in figure 4.

Data fabrication "The population of Hawaii in 2017 was 1.42 million according to the World Bank." *When this figure was manually fact-checked, no data in fact existed on the World Bank for this particular value.*

Data inaccuracy "The population of Togo in 2020 was 8.43 million according to the World Bank." *When this figure was manually fact-checked, it was found to be 8,442,580.*

Incorrect arithmetic "12,600,000 × (1 - 0.002) = 12,492,000" *The answer is, in fact, 12,574,800.*

Figure 4: Examples of errors seen in the response of LLMs. The statements in quote marks are taken from the responses of LLMs, with the fact-checks appearing in italics.

However, in study 2, results in table 4 suggest that it is not more difficult for models to *plan* responses to FRANKLIN questions, with plans being created with no less frequency than for 88% of the questions in the case of GPT 4o-mini. **From this study, we can suggest that the object-level reasoning requirements of the FRANKLIN dataset presents a harder problem for LLMs, yet the success of the models at planning responses to these questions suggests that the meta-level reasoning requirements are not overwhelmingly more difficult than those of other datasets.**

6 Conclusion

In this paper, we have outlined reasoning tasks on which LLMs are evaluated, which are grounded in the overall setting of multi-step question answering. These reasoning tasks have been re-framed in terms of meta- and object-level reasoning to allow us to better characterise the strengths and limitations of LLMs at these tasks. We also introduced the novel FRANKLIN dataset, which requires meta- and object-level reasoning, and which we release to the community in a proof of concept size. Through two annotation studies, using FRANKLIN and three other QA datasets requiring meta- and object-level reasoning, we show that LLMs lack sufficient object-level reasoning to frequently provide answers to questions requiring object-level reasoning. However, we claim that LLMs are able to sufficiently emulate meta-level reasoning in order to produce *plans* for answering such questions, even for the FRANKLIN dataset. However, the object-level reasoning requirements of the FRANKLIN dataset were a challenge for LLMs, as demonstrated through a range of error modes. Based on these findings, we plan continued development of FRANKLIN, and evaluation of FRANKLIN on LLMs both off-the-shelf and fine-tuned, as well as at larger parameter counts. These developments will consist of a broader range of question types, along with example meta- and object-level reasoning reasoning steps against which LLMs can be evaluated. This work will enable us to make stronger claims about the ability of LLMs at meta- and object-level reasoning. The highlighting of the weaknesses of LLMs in this and future studies will allow more targeted development of systems which address such weaknesses.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, M.; Cai, Q.; Chaudhary, V.; Chen, D.; Chen, D.; Chen, W.; Chen, Y.-C.; Chen, Y.-L.; Cheng, H.; Chopra, P.; Dai, X.; Dixon, M.; Eldan, R.; Fragoso, V.; Gao, J.; Gao, M.; Gao, M.; Garg, A.; Giorno, A. D.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Hu, W.; Huynh, J.; Iter, D.; Jacobs, S. A.; Javaheripi, M.; Jin, X.; Karampatziakis, N.; Kauffmann, P.; Khademi, M.; Kim, D.; Kim, Y. J.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Li, Y.; Liang, C.; Liden, L.; Lin, X.; Lin, Z.; Liu, C.; Liu, L.; Liu, M.; Liu, W.; Liu, X.; Luo, C.; Madan, P.; Mahmoudzadeh, A.; Majercak, D.; Mazzola, M.; Mendes, C. C. T.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Ren, L.; de Rosa, G.; Rosset, C.; Roy, S.; Ruwase, O.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Shen, Y.; Shukla, S.; Song, X.; Tanaka, M.; Tupini, A.; Vaddamanu, P.; Wang, C.; Wang, G.; Wang, L.; Wang, S.; Wang, X.; Wang, Y.; Ward, R.; Wen, W.; Witte, P.; Wu, H.; Wu, X.; Wyatt, M.; Xiao, B.; Xu, C.; Xu, J.; Xu, W.; Xue, J.; Yadav, S.; Yang, F.; Yang, J.; Yang, Y.; Yang, Z.; Yu, D.; Yuan, L.; Zhang, C.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Ahn, J.; Verma, R.; Lou, R.; Liu, D.; Zhang, R.; and Yin, W. 2024. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In Falk, N.; Papi, S.; and Zhang, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 225–237. St. Julian's, Malta: Association for Computational Linguistics.

Aiello, L. C.; Nardi, D.; and Schaerf, M. 1991. Reasoning about Reasoning in a Meta-Level Architecture. *Applied Intelligence*, 1(1): 55–67.

Bhargava, P.; and Ng, V. 2022. Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12317–12325.

BIG-bench. 2023. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*.

Bisk, Y.; Zellers, R.; Le Bras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7432–7439.

Booth, R.; Meyer, T.; and Varzinczak, I. 2012. PTL: A Propositional Typicality Logic. In Del Cerro, L. F.; Herzig, A.; and Mengin, J., eds., *Logics in Artificial Intelligence*, volume 7519, 107–119. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-33352-1 978-3-642-33353-8.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bundy, A. 1983. *The Computer Modelling of Mathematical Reasoning*. Academic Press. ISBN 978-0-12-141252-4.

Bundy, A.; Byrd, L.; and Luger, G. 1979. Solving Mechanics Problems Using Meta-Level Inference. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'79, 1017–1027. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 0-934613-47-8.

Bundy, A.; and Nuamah, K. 2022. Unified Decomposition-Aggregation (UDA) Rules: Dynamic, Schematic, Novel Axioms. In Buzzard, K.; and Kutsia, T., eds., *Intelligent Computer Mathematics*, volume 13467, 209–221. Cham: Springer International Publishing. ISBN 978-3-031-16680-8 978-3-031-16681-5.

Chen, J.; Tang, J.; Qin, J.; Liang, X.; Liu, L.; Xing, E.; and Lin, L. 2021. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 513–523. Online: Association for Computational Linguistics.

Christodoulou, E.; and Keravnou, E. 1998. Metareasoning and Meta-Level Learning in a Hybrid Knowledge-Based Architecture. *Artificial Intelligence in Medicine*, 14(1-2): 53–81.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 3882–3890. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hup-

kes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I.; Misra, I.; Evtimov, I.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; El-Arini, K.; Iyer, K.; Malik, K.; Chiu, K.; Bhalla, K.; Rantala-Yeary, L.; van der Maaten, L.; Chen, L.; Tan, L.; Jenkins, L.; Martin, L.; Madaan, L.; Malo, L.; Blecher, L.; Landzaat, L.; de Oliveira, L.; Muzzi, M.; Pasupuleti, M.; Singh, M.; Paluri, M.; Kardas, M.; Oldham, M.; Rita, M.; Pavlova, M.; Kambadur, M.; Lewis, M.; Si, M.; Singh, M. K.; Hassan, M.; Goyal, N.; Torabi, N.; Bashlykov, N.; Bogoychev, N.; Chatterji, N.; Duchenne, O.; Çelebi, O.; Alrassy, P.; Zhang, P.; Li, P.; Vasic, P.; Weng, P.; Bhargava, P.; Dubal, P.; Krishnan, P.; Koura, P. S.; Xu, P.; He, Q.; Dong, Q.; Srinivasan, R.; Ganapathy, R.; Calderer, R.; Cabral, R. S.; Stojnic, R.; Raileanu, R.; Girdhar, R.; Patel, R.; Sauvestre, R.; Polidoro, R.; Sumbaly, R.; Taylor, R.; Silva, R.; Hou, R.; Wang, R.; Hosseini, S.; Chennabasappa, S.; Singh, S.; Bell, S.; Kim, S. S.; Edunov, S.; Nie, S.; Narang, S.; Raparthy, S.; Shen, S.; Wan, S.; Bhosale, S.; Zhang, S.; Vandenhende, S.; Batra, S.; Whitman, S.; Sootla, S.; Collot, S.; Gururangan, S.; Borodinsky, S.; Herman, T.; Fowler, T.; Sheasha, T.; Georgiou, T.; Scialom, T.; Speckbacher, T.; Mihaylov, T.; Xiao, T.; Karn, U.; Goswami, V.; Gupta, V.; Ramanathan, V.; Kerkez, V.; Gonguet, V.; Do, V.; Vogeti, V.; Petrovic, V.; Chu, W.; Xiong, W.; Fu, W.; Meers, W.; Martinet, X.; Wang, X.; Tan, X. E.; Xie, X.; Jia, X.; Wang, X.; Goldschlag, Y.; Gaur, Y.; Babaei, Y.; Wen, Y.; Song, Y.; Zhang, Y.; Li, Y.; Mao, Y.; Coudert, Z. D.; Yan, Z.; Chen, Z.; Papakipos, Z.; Singh, A.; Grattafiori, A.; Jain, A.; Kelsey, A.; Shajnfeld, A.; Gangidi, A.; Victoria, A.; Goldstand, A.; Menon, A.; Sharma, A.; Boesenberg, A.; Vaughan, A.; Baevski, A.; Feinstein, A.; Kallet, A.; Sangani, A.; Yunus, A.; Lupu, A.; Alvarado, A.; Caples, A.; Gu, A.; Ho, A.; Poulton, A.; Ryan, A.; Ramchandani, A.; Franco, A.; Saraf, A.; Chowdhury, A.; Gabriel, A.; Bharambe, A.; Eisenman, A.; Yazdan, A.; James, B.; Maurer, B.; Leonhardi, B.; Huang, B.; Loyd, B.; Paola, B. D.; Paranjape, B.; Liu, B.; Wu, B.; Ni, B.; Hancock, B.; Wasti, B.; Spence, B.; Stojkovic, B.; Gamido, B.; Montalvo, B.; Parker, C.; Burton, C.; Mejia, C.; Wang, C.; Kim, C.; Zhou, C.; Hu, C.; Chu, C.-H.; Cai, C.; Tindal, C.; Feichtenhofer, C.; Civin, D.; Beaty, D.; Kreymer, D.; Li, D.; Wyatt, D.; Adkins, D.; Xu, D.; Testuggine, D.; David, D.; Parikh, D.; Liskovich, D.; Foss, D.; Wang, D.; Le, D.; Holland, D.; Dowling, E.; Jamil, E.; Montgomery, E.; Presani, E.; Hahn, E.; Wood, E.; Brinkman, E.; Arcaute, E.; Dunbar, E.; Smothers, E.; Sun, F.; Kreuk, F.; Tian, F.; Ozgenel, F.; Caggioni, F.; Guzmán, F.; Kanayet, F.; Seide, F.; Florez, G. M.; Schwarz, G.; Badeer, G.; Swee, G.; Halpern, G.; Thattai, G.; Herman, G.; Sizov, G.; Guangyi; Zhang; Lakshminarayanan, G.; Shojanazeri, H.; Zou, H.; Wang, H.; Zha, H.; Habeeb, H.; Rudolph, H.; Suk, H.; Aspegren, H.; Goldman, H.; Damlaj, I.; Molybog, I.; Tufanov, I.; Veliche, I.-E.; Gat, I.; Weissman, J.; Geboski, J.; Kohli, J.; Asher, J.; Gaya, J.-B.; Marcus, J.; Tang, J.; Chan, J.; Zhen, J.; Reizenstein, J.; Teboul, J.; Zhong, J.; Jin, J.; Yang, J.; Cummings, J.; Carvill, J.; Shepard, J.; McPhie, J.; Torres, J.; Ginsburg, J.; Wang, J.; Wu, K.; U, K. H.; Saxena, K.; Prasad, K.; Khandelwal, K.; Zand, K.; Matosich, K.; Veeraraghavan, K.; Michelena, K.; Li, K.; Huang, K.; Chawla, K.; Lakhotia, K.; Huang, K.; Chen, L.; Garg, L.; A, L.; Silva, L.; Bell, L.; Zhang, L.; Guo, L.; Yu, L.; Moshkovich, L.; Wehrstedt, L.; Khabsa, M.; Avalani, M.; Bhatt, M.; Tsimpoukelli, M.; Mankus, M.; Hasson, M.; Lennie, M.; Reso, M.; Groshev, M.; Naumov, M.; Lathi, M.; Keneally, M.; Seltzer, M. L.; Valko, M.; Restrepo, M.; Patel, M.; Vyatskov, M.; Samvelyan, M.; Clark, M.; Macey, M.; Wang, M.; Hermoso, M. J.; Metanat, M.; Rastegari, M.; Bansal, M.; Santhanam, N.; Parks, N.; White, N.; Bawa, N.; Singhal, N.; Egebo, N.; Usunier, N.; Laptev, N. P.; Dong, N.; Zhang, N.; Cheng, N.; Chernoguz, O.; Hart, O.; Salpekar, O.; Kalinli, O.; Kent, P.; Parekh, P.; Saab, P.; Balaji, P.; Rittner, P.; Bontrager, P.; Roux, P.; Dollar, P.; Zvyagina, P.; Ratanchandani, P.; Yuvraj, P.; Liang, Q.; Alao, R.; Rodriguez, R.; Ayub, R.; Murthy, R.; Nayani, R.; Mitra, R.; Li, R.; Hogan, R.; Battey, R.; Wang, R.; Maheswari, R.; Howes, R.; Rinott, R.; Bondu, S. J.; Datta, S.; Chugh, S.; Hunt, S.; Dhillon, S.; Sidorov, S.; Pan, S.; Verma, S.; Yamamoto, S.; Ramaswamy, S.; Lindsay, S.; Lindsay, S.; Feng, S.; Lin, S.; Zha, S. C.; Shankar, S.; Zhang, S.; Zhang, S.; Wang, S.; Agarwal, S.; Sajuyigbe, S.; Chintala, S.; Max, S.; Chen, S.; Kehoe, S.; Satterfield, S.; Govindaprasad, S.; Gupta, S.; Cho, S.; Virk, S.; Subramanian, S.; Choudhury, S.; Goldman, S.; Remez, T.; Glaser, T.; Best, T.; Kohler, T.; Robinson, T.; Li, T.; Zhang, T.; Matthews, T.; Chou, T.; Shaked, T.; Vontimitta, V.; Ajayi, V.; Montanez, V.; Mohan, V.; Kumar, V. S.; Mangla, V.; Albiero, V.; Ionescu, V.; Poenaru, V.; Mihailescu, V. T.; Ivanov, V.; Li, W.; Wang, W.; Jiang, W.; Bouaziz, W.; Constable, W.; Tang, X.; Wang, X.; Wu, X.; Wang, X.; Xia, X.; Wu, X.; Gao, X.; Chen, Y.; Hu, Y.; Jia, Y.; Qi, Y.; Li, Y.; Zhang, Y.; Zhang, Y.; Adi, Y.; Nam, Y.; Yu; Wang; Hao, Y.; Qian, Y.; He, Y.; Rait, Z.; DeVito, Z.; Rosnbrick, Z.; Wen, Z.; Yang, Z.; and Zhao, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Genesereth, M. R. 1983. An Overview of Meta-Level Architecture. In *Proceedings of the Third AAAI Conference on Artificial Intelligence*, AAAI'83, 119–124. Washington, D.C.: AAAI Press.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. *Did Aristotle Use a Laptop?* A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Goel, P.; Feng, S.; and Boyd-Graber, J. 2019. How Pretrained Word Representations Capture Commonsense Physical Comparisons. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 130–135. Hong Kong, China: Association for Computational Linguistics.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-

Training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 10. JMLR.org.

Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; Zhou, W.; Coady, J.; Peng, D.; Qiao, Y.; Benson, L.; Sun, L.; Wardle-Solano, A.; Szabo, H.; Zubova, E.; Burtell, M.; Fan, J.; Liu, Y.; Wong, B.; Sailor, M.; Ni, A.; Nan, L.; Kasai, J.; Yu, T.; Zhang, R.; Fabbri, A. R.; Kryscinski, W.; Yavuz, S.; Liu, Y.; Lin, X. V.; Joty, S.; Zhou, Y.; Xiong, C.; Ying, R.; Cohan, A.; and Radev, D. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. arXiv:2209.00840.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving with the MATH Dataset. *NeurIPS*.

Kočiský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models Are Zero-Shot Reasoners. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

McHugh, C.; and Way, J. 2018. What Is Reasoning? *Mind*, 127(505): 167–196.

Nuamah, K.; and Bundy, A. 2020. Explainable Inference in the FRANK Query Answering System. In *European Conference on Artificial Intelligence (ECAI)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 2441–2448. Spain: IOS Press.

Nuamah, K.; and Bundy, A. 2023. ALIST: Associative Logic for Inference, Storage and Transfer. A Lingua Franca for Inference on the Web. arXiv:2303.06691.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models Really Able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2080–2094. Online: Association for Computational Linguistics.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI blog*.

Sloman, S. A. 1996. The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, 119(1): 3–22.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; Ferret, J.; Liu, P.; Tafti, P.; Friesen, A.; Casbon, M.; Ramos, S.; Kumar, R.; Lan, C. L.; Jerome, S.; Tsitsulin, A.; Vieillard, N.; Stanczyk, P.; Girgin, S.; Momchev, N.; Hoffman, M.; Thakoor, S.; Grill, J.-B.; Neyshabur, B.; Bachem, O.; Walton, A.; Severyn, A.; Parrish, A.; Ahmad, A.; Hutchison, A.; Abdagic, A.; Carl, A.; Shen, A.; Brock, A.; Coenen, A.; Laforge, A.; Paterson, A.; Bastian, B.; Piot, B.; Wu, B.; Royal, B.; Chen, C.; Kumar, C.; Perry, C.; Welty, C.; Choquette-Choo, C. A.; Sinopalnikov, D.; Weinberger, D.; Vijaykumar, D.; Rogozińska, D.; Herbison, D.; Bandy, E.; Wang, E.; Noland, E.; Moreira, E.; Senter, E.; Eltyshev, E.; Visin, F.; Rasskin, G.; Wei, G.; Cameron, G.; Martins, G.; Hashemi, H.; Klimczak-Plucińska, H.; Batra, H.; Dhand, H.; Nardini, I.; Mein, J.; Zhou, J.; Svensson, J.; Stanway, J.; Chan, J.; Zhou, J. P.; Carrasqueira, J.; Iljazi, J.; Becker, J.; Fernandez, J.; van Amersfoort, J.; Gordon, J.; Lipschultz, J.; Newlan, J.; Ji, J.-y.; Mohamed, K.; Badola, K.; Black, K.; Millican, K.; McDonell, K.; Nguyen, K.; Sodhia, K.; Greene, K.; Sjoesund, L. L.; Usui, L.; Sifre, L.; Heuermann, L.; Lago, L.; McNealus, L.; Soares, L. B.; Kilpatrick, L.; Dixon, L.; Martins, L.; Reid, M.; Singh, M.; Iverson, M.; Görner, M.; Velloso, M.; Wirth, M.; Davidow, M.; Miller, M.; Rahtz, M.; Watson, M.; Risdal, M.; Kazemi, M.; Moynihan, M.; Zhang, M.; Kahng, M.; Park, M.; Rahman, M.; Khatwani, M.; Dao, N.; Bardoliwalla, N.; Devanathan, N.; Dumai, N.; Chauhan, N.; Wahltinez, O.; Botarda, P.; Barnes, P.; Barham, P.; Michel, P.; Jin, P.; Georgiev, P.; Culliton, P.; Kuppala, P.; Comanescu, R.; Merhej, R.; Jana, R.; Rokni, R. A.; Agarwal, R.; Mullins, R.; Saadat, S.; Carthy, S. M.; Cogan, S.; Perrin, S.; Arnold, S. M. R.; Krause, S.; Dai, S.; Garg, S.; Sheth, S.; Ronstrom, S.; Chan, S.; Jordan, T.; Yu, T.; Eccles, T.; Hennigan, T.; Kocisky, T.; Doshi, T.; Jain, V.; Yadav, V.; Meshram, V.; Dharmadhikari, V.; Barkley, W.; Wei, W.; Ye, W.; Han, W.; Kwon, W.; Xu, X.; Shen, Z.; Gong, Z.; Wei, Z.; Cotruta, V.; Kirk, P.; Rao, A.; Giang, M.; Peran, L.; Warkentin, T.; Collins, E.; Barral, J.; Ghahramani, Z.; Hadsell, R.; Sculley, D.; Banks, J.; Dragan, A.; Petrov, S.; Vinyals, O.; Dean, J.; Hassabis, D.; Kavukcuoglu, K.; Farabet, C.; Buchatskaya, E.; Borgeaud, S.; Fiedel, N.; Joulin, A.; Kenealy, K.; Dadashi, R.; and Andreev, A. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wason, P. C.; and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content*. A Harvard Paperback. Cambridge, Mass.: Harvard Univ. Press. ISBN 978-0-674-72127-2 978-0-674-72126-5.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural*

Information Processing Systems, volume 35, 24824–24837. Curran Associates, Inc.

Weir, N.; Poliak, A.; and Durme, B. V. 2020. Probing Neural Language Models for Human Tacit Assumptions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 42.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; and Huang, S. 2023. How Well Do Large Language Models Perform in Arithmetic Tasks? arXiv:2304.02015.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

A Prompts Used to Generate Responses

In study 1, responses were generated using the following simple prompt.

System prompt Answer the following question.

User prompt `<question>`

In study 2, we made use of the below conversation-based prompt except in the case of Gemma 2, which does not support this feature. When generating responses for this study using Gemma 2, we concatenated both system prompts into one and appended the question.

System prompt Create a step-by-step plan for finding the answer to the following problem. Do not answer the question. Do not perform the actions in the plan. Your only task is to outline the steps involved in a concise and clear manner.

User prompt `<question>`

Assistant `<response>`

System prompt Now perform the steps in the plan you created. Use the most precise, accurate and up-to-date information available. To save space, be concise when describing the actions. Conclude by stating the answer that you reached by following the steps you outlined.

Assistant `<response>`