MULTI-TASK INFERENCE: Can Large Language Models Follow Multiple Instructions at Once?

Guijin Son[1,2,3] **Sangwon Baek**[2] **Sangdae Nam**[2,5] **Ilgyun Jeong**[2,4] **Seungone Kim**[6,7*]

Yonsei University[1] EleutherAI[2] OneLineAI[3] Korea University[4]
VESSL AI[5] KAIST AI[6] Carnegie Mellon University [7]
spthsrbwls123@yonsei.ac.kr seungone@kaist.ac.kr

Abstract

Large language models (LLMs) are typically prompted to follow a *single* instruction per inference call. In this work, we analyze whether LLMs also hold the capability to handle *multiple* instructions simultaneously, denoted as MULTI-TASK INFERENCE. For this purpose, we introduce the MTI BENCH (**M**ulti-**T**ask **I**nference **Bench**mark), a comprehensive evaluation benchmark encompassing 5,000 instances across 25 tasks. Each task in the MTI BENCH involves 2 to 3 sub-tasks. As expected, we first demonstrate that MULTI-TASK INFERENCE reduces the total inference time by $\times 1.46$ times in average since it does not require multiple inference calls. Interestingly, contrary to the expectation that LLMs would perform better when tasks are divided, we find that state-of-the-art LLMs, such as LLAMA-2-CHAT-70B and GPT-4, show up to 7.3% and 12.4% improved performance with MULTI-TASK INFERENCE compared to SINGLE-TASK INFERENCE on the MTI BENCH. We release the MTI BENCH dataset and our code at this link[1].

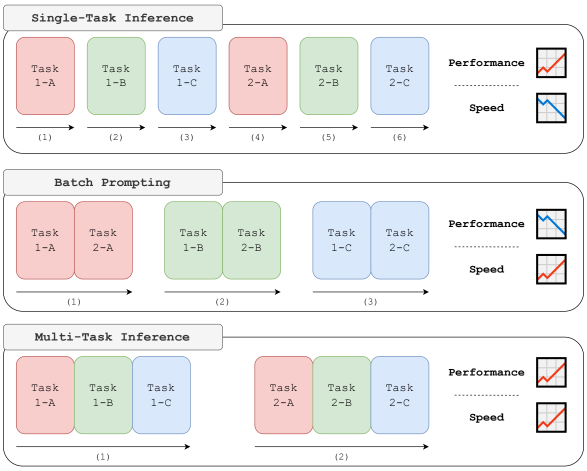

Figure 1: Comparison of the three inference methods for handling tasks composed of three sub-tasks: SINGLE-TASK INFERENCE, BATCH PROMPTING, and MULTI-TASK INFERENCE. MULTI-TASK INFERENCE shows reliable performance as SINGLE-TASK INFERENCE and provides faster speed as BATCH PROMPTING (Cheng et al., 2023).

1 Introduction

Large language models (LLMs) capable of following instructions have demonstrated impressive performance across a wide range of tasks (Xu et al., 2023; OpenAI, 2023; Anil et al., 2023; Tunstall et al., 2023; Wang et al., 2023a). However, since LLMs are trained to follow a *single* instruction per inference call, it is questionable whether they also hold the ability to follow complex instructions that necessitate handling *multiple* sub-tasks (Yang et al., 2018; Geva et al., 2021; Cheng et al., 2023). Moreover, current evaluation resources are either confined to measuring the LLM's capability in following one-step instructions (Li et al., 2023; Chiang et al., 2023; Zheng et al., 2023) or only diagnose the capability to process multi-step instructions in a particular domain such as commonsense reasoning and arithmetic (Geva et al., 2021; Cobbe et al., 2021; Lightman et al., 2023).

In this paper, we analyze whether LLMs hold the capability to handle tasks composed of multiple instructions at one inference call, which we denote as MULTI-TASK INFERENCE. As shown in Figure 1, we compare the performance and speed with two baselines: (1) SINGLE-TASK INFERENCE: addressing sub-tasks sequentially and (2) BATCH PROMPTING: simultaneously processing multiple instances from the same task (Cheng et al., 2023).

For this purpose, we construct the MTI BENCH (**M**ulti-**T**ask **I**nference **Bench**mark), an evaluation dataset featuring 25 tasks, each consisting of 2 to 3 sub-tasks. As shown in Figure 2, the MTI BENCH is divided into two distinct subsets: (a) the MULTI-STEP subset, which evaluates the models' ability follow multiple instructions sequentially and (b) the MULTI-PART subset, focusing on the models'

[*]Corresponding Author
[1]https://github.com/guijinSON/MTI-Bench

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5606–5627
August 11-16, 2024

capability to handle multiple sub-tasks that do not have a sequential dependency. Notably, the MTI BENCH sets itself apart from previous multi-hop reasoning (Yang et al., 2018; Geva et al., 2021) and multi-turn conversation (Zheng et al., 2023) evaluation suites by providing annotations to assess the *intermediate* performance of LLMs while solving multi-task instructions. This enables researchers to check if LLMs reach the correct answers and evaluate whether their reasoning process is consistent and logical throughout the process.

We evaluate 11 LLMs capable of following instructions, varying in parameter size. Surprisingly, on the MTI BENCH, state-of-the-art LLMs such as LLAMA-2-CHAT-70B and GPT-4 show up to 7.3% and 12.4% better performance with MULTI-TASK INFERENCE compared to SINGLE-TASK INFERENCE. Moreover, MULTI-TASK INFERENCE requires x1.46 times less average inference time than SINGLE-TASK INFERENCE. These results indicate that users could obtain similar performance with substantially less time when querying instructions that necessitate solving multiple sub-tasks. Through ablation experiments, we suggest that looking at the next sub-task provides critical clues on the answer format for solving the previous sub-task.

Our contributions are as follows:

- We are, to the best of our knowledge, the first to develop an evaluation benchmark, the MTI BENCH, tailored to analyze the MULTI-TASK INFERENCE capabilities of LLMs. We fully open-source our code and data.
- Our findings demonstrate that MULTI-TASK INFERENCE surprisingly works well compared to SINGLE-TASK INFERENCE only for stronger models.
- We show that MULTI-TASK INFERENCE offers x1.46 times speed-up compared to SINGLE-TASK INFERENCE. This suggests that practitioners can fully leverage the capability of LLMs to solve multiple tasks at one inference call.

2 Related Works

2.1 Language Model Evaluation

While Large Language Models (LLMs) demonstrate impressive performance across a wide range of tasks, it remains essential to assess their properties and behaviors from various perspectives (Chang et al., 2023; Wang et al., 2023b; Chia et al., 2023). Traditionally, evaluations of LLMs primarily focused on performance in specific domains or tasks (Hendrycks et al., 2020; Srivastava et al., 2023). However, there is a growing interest in holistically evaluating LLMs' properties and high-level capabilities across multiple facets (Liang et al., 2022; Holtzman et al., 2023; Kim et al., 2023b). Prior research in this area includes measuring overall helpfulness and harmlessness in user interactions (Dubois et al., 2023; Li et al., 2023; Zheng et al., 2023), assessing the ability to generate coherent thought chains in reasoning tasks (Fu et al., 2023; Ott et al., 2023), examining the presence of a theory of mind (Zhou et al., 2023; Kim et al., 2023a; Mireshghallah et al., 2023), and evaluating the capacity to avoid producing toxic content (Gehman et al., 2020). In our work, we focus on multi-processing capabilities, specifically the ability of LLMs to process multiple instructions simultaneously, as a novel and significant area to explore and evaluate across various LLMs.

2.2 Multiprocessing Capabilities of LLMs

The ability to concurrently process multiple pieces of information is a key indicator of intelligence (Meyer and Kieras, 1997). Previous studies have introduced datasets like HotpotQA (Yang et al., 2018) and StrategyQA (Geva et al., 2021), which require multi-hop reasoning. These are designed to train and test the LLM's capability to follow the internal reasoning processes needed for a valid final prediction. However, these datasets do not offer a comprehensive method to assess the accuracy of intermediate steps or to compare concurrent versus sequential processing. Recently, Cheng et al. (2023) introduced BATCH PROMPTING, aligning with the research direction of our study. However, this approach is limited to examining if LLMs can process multiple instances within the *same* task. In contrast, our MTI BENCH encompasses a broader range of scenarios, including instructions comprising multiple sub-tasks that either follow a sequential order (MULTI-STEP subset) or solve different tasks (MULTI-PART).

3 The MTI BENCH Dataset

In this section, we explain how the MTI BENCH is formulated (Section 3.1), how we constructed it (Section 3.2), and provide an analysis of the diver-

	# of		Avg. Length	
	Task	Task Type	Instruction	Context
MULTI-STEP	13	12	20.3	89.4
MULTI-PART	12	16	22.4	104.8
TOTAL	25	28	17.4	115.8

Table 1: Dataset Statistics for MTI BENCH. The lengths of instructions and context are measured in the number of words.

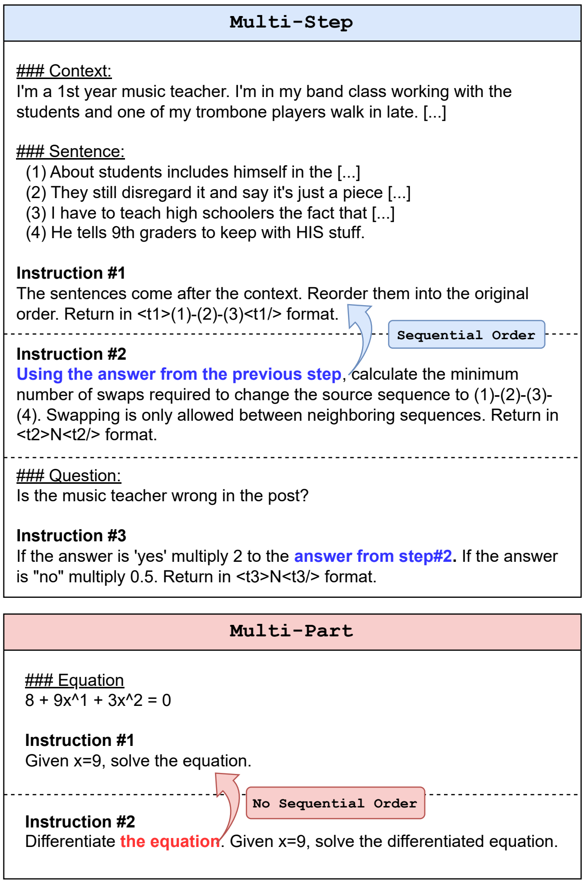

Figure 2: An example comparing the MULTI-PART and MULTI-STEP subset within the MTI BENCH dataset. Whereas the MULTI-STEP necessitates to solve step-by-step since there is a sequential order among the sub-tasks, the sub-tasks within the MULTI-PART does not have a sequential order.

sity, compositionality, and quality. (Section 3.3).

3.1 Task Formulation

The MTI BENCH (**M**ulti-**T**ask **I**nference **Bench**mark) is a comprehensive benchmark to evaluate the MULTI-TASK INFERENCE capabilities of LLMs. The benchmark comprises 25 tasks, each with 200 instances, summing up to 5,000 instances in total. Each task within the benchmark comprises 2 to 3 sub-tasks, selected from a diverse pool of 28 NLP tasks, including Classification, Multiple-Choice Question Answering (MCQA), Arithmetic, and Natural Language Inference. These tasks are divided into two subsets: MULTI-STEP and MULTI-PART containing 13 and 12 tasks respectively. Five of the 25 tasks consist of 3 sub-tasks. Table 1 presents detailed statistics for each subset.

Tasks in the MULTI-STEP subset demand a sequential approach, with the accuracy of each step being vital for the following ones. This subset assesses LMs' proficiency in managing interdependent tasks. Conversely, the MULTI-PART subset consists of contextually related but independent sub-tasks, evaluating LLMs' capacity to process multiple, disparate tasks simultaneously. Both subsets employ exact string matching as the evaluation method, focusing on both intermediate and final accuracy. An example instance for each subset is illustrated in Figure 2.

3.2 Dataset Construction

To construct the MTI BENCH, we select a wide range of tasks from existing NLP benchmarks. Our primary sources include Quoref (Dasigi et al., 2019), SNLI (Bowman et al., 2015a), MMLU (Hendrycks et al., 2020), and MATH (Hendrycks et al., 2021). Tables 13-37 provides a comprehensive list of datasets used to construct the benchmark. The key criteria for source dataset selection are (1) the presence of a rigorous quality control process in the datasets and (2) the potential to integrate the datasets into more complex tasks. The co-authors split into two groups for efficiency: one focused on combining different tasks into composite tasks, while the other screened for and eliminated any combinations that were uninformative or of low quality, subsequently categorizing the tasks into either MULTI-STEP, or MULTI-PART subsets. During the process, 7 out of the initial 32 multi-tasks were deemed unsuitable and removed, resulting in a refined final version of 25 high-quality tasks. Additionally, we crafted a one-shot demonstration for each task, which sequentially resolves the sub-tasks by generating a Chain-of-Thought (Wei et al., 2022b).

3.3 Dataset Analysis

Diversity The distribution of NLP tasks in their respective order within the sub-tasks is detailed in Table 2. No single task type dominates, ensuring a wide-ranging evaluation of model capabilities. There are only five multi-tasks comprised of three sub-tasks, resulting in a relatively constrained diversity for 3RD sub-task.

Task Type	1ST	2ND	3RD
Others	32%	24%	-
Classification	28%	4%	-
Sentence Sorting	20%	12%	-
Answerability Classification	16%	4%	-
Natural Language Inference	4%	8%	-
Extractive QA	-	4%	40%
Arithmetic	-	16%	20%
Multiple-Choice QA	-	12%	20%
Binary QA	-	4%	20%
Wrong Candidate Ranking	-	8%	-
Judicial Decision	-	4%	-

Table 2: Distribution of Task Types for each sub-task.

Subset	Chi-square Statistic	p-value	Odds Ratio
MULTI-STEP	394.37	< 0.0001	8.44
MULTI-PART	128.28	< 0.0001	2.64

Table 3: Chi-squared statistical test results for MULTI-STEP and MULTI-PART subsets. The results indicate that the tasks are properly classified into each category as intended.

Compositionality To statistically verify the authors' manual classification of multi-tasks into MULTI-STEP and MULTI-PART subset, we conduct a chi-squared test to study the interdependency within each subset. Initially, a GPT-3.5-TURBO model was used to solve 200 instances of each multi-task combination. Subsequently, a chi-squared test was applied to the outcomes to assess the dependency between the accuracy of each sub-task. In Table 3, both subsets demonstrated p-values below the 0.01 threshold, refuting the null hypothesis that the sub-tasks are independent. Furthermore, the MULTI-STEP subset features chi-square statistic and odds ratio substantially higher than the MULTI-PART subset, indicating a more pronounced linear association among its tasks.

Quality To ensure the quality of the MTI BENCH, we conduct a two-step quality check. Initially, we selected a random sample of eight instructions from each task, making a total of 200 instructions for evaluation. Two of our authors labeled whether each instance showed valid dependencies between sub-tasks and were properly categorized. Tasks were recategorized and rephrased according to the results. After these adjustments, a final round of quality assessment was conducted. This phase involved ten professional annotators, including authors from our team and five externally recruited experts. The hired experts, all master's graduates in finance, business, and computer science, were paid at the rate of $0.11 per question.

Quality Review Question	1ST	FINAL
Does the instruction feature valid sub-task dependencies?	89%	91%
Is the (instruction, context, answer) triplet suitable for the benchmark?	88%	92%
Does the task align with its designated category (MULTI-STEP, MULTI-PART)?	76%	88%
All fields are invalid	1%	0%

Table 4: Data quality review for each component within the MTI BENCH instance: the instruction, context, answer. Annotators were asked to answer either "Yes" or "No" for each question given a randomly sampled instance from the MTI Bench. Results show the ratio of "Yes" from the annotators.

The evaluation results, presented in Table 4, indicate that after the modification process, majority of the multi-tasks in the benchmark demonstrate valid sub-task dependencies and are correctly categorized. Two annotators reviewed each question, and the Cohen's kappa statistic (McHugh, 2012) for inter-annotator agreement on these questions scored 0.82, 0.68, and 0.89, indicating a substantial level of consensus. It was also noted that the remaining misclassifications did not reflect the overall task labeling but were somewhat isolated incidents, likely due to the specific contexts of individual samples. Importantly, even in cases with errors, no instances fail the quality assessment criteria completely, suggesting that the errors were not severe enough to affect the dataset's reliability as a benchmarking tool.

4 Experimental Setup

In this section, we explain our experimental setup for investigating the MULTI-TASK INFERENCE capabilities of LLMs.

Baseline Inference Methods In addition to MULTI-TASK INFERENCE, the method in our main consideration, we compare with SINGLE-TASK INFERENCE and BATCH PROMPTING (Cheng et al., 2023). Figure 1 illustrates a scenario that compares the three inference methods. Assuming that we are testing an LLM with two instances that consist of 3 sub-tasks, the most naive approach, SINGLE-TASK INFERENCE prompts an LLM 6 times, where each inference call corresponds to solve a single sub-task. On the other hand, BATCH PROMPTING groups the same sub-tasks and prompts an LLM to solve multiple instances at once. Lastly, MULTI-TASK INFERENCE prompts the LLM to solve all

	Single-Task			Batch Prompting			Multi-Task		
	M.S.	M.P.	Average	M.S.	M.P.	Average	M.S.	M.P.	Average
Tulu-7B	0.9	0.9	0.9	0.0	0.0	0.0	0.4	1.5	**1.0**
Tulu-13B	2.9	2.6	2.8	0.0	0.0	0.0	2.1	3.8	**3.0**
Tulu-30B	8.2	5.4	**6.8**	2.0	1.0	1.5	1.5	4.4	3.0
Tulu-65B	1.4	4.6	3.0	2.4	3.0	2.7	5.6	7.1	**6.4**
Llama-2-Chat-7B	2.8	4.4	3.6	1.0	0.0	0.5	5.5	7.9	**6.7**
Llama-2-Chat-13B	1.0	3.0	2.0	0.0	0.0	0.0	2.4	4.2	**3.3**
Llama-2-Chat-70B	8.0	9.4	8.7	7.4	8.3	7.9	16.0	20.0	**18.0**
Vicuna-7B	2.2	2.3	2.3	1.3	1.5	1.4	3.9	4.8	**4.4**
Vicuna-13B	6.5	11.6	**9.1**	2.4	1.9	2.2	7.3	9.3	8.3
GPT-3.5-Turbo	18.9	23.7	21.3	18.1	19.1	18.6	21.5	26.2	**23.9**
GPT-4	25.8	35.7	30.8	33.3	31.0	32.2	43.2	42.5	**42.9**

Table 5: Evaluation results of Multi-step (M.S.), and multi-part (M.P.) subset utilizing single-task inference, batch-prompting and multi-task inference. The specified accuracy is the accuracy of correctly completing all sub-tasks (i.e., **final accuracy**). Evaluations are held in a one-shot setting with chain-of-thought reasoning. The best comparable performances among the inference methods are bolded and underlined.

the multiple sub-tasks within a single inference call. In general, if N instances consisting of M sub-tasks are given, Single-Task Inference requires N times more inference calls compared to Batch Prompting and M times more inference calls compared to Multi-Task Inference.

Test Models We evaluate eleven LLMs capable of following instructions including: (1) GPT-4 (OpenAI, 2023), (2) GPT-3.5 (OpenAI, 2022), (3) Tulu (7b, 13b, 30b, 65b) (Wang et al., 2023c), (4) vicuna (7b, 13b) (Chiang et al., 2023), and (5) Llama-2-Chat (7b, 13b, 70b) (Touvron et al., 2023). For GPT-4 and GPT-3.5, we utilize the 0613 version. Reported results represent the average of three runs, except for GPT-4, which were evaluated in a single run to minimize costs. Open-source models were run using fp16 precision. All evaluations were conducted in a single-shot setting, incorporating Chain-of-Thought reasoning. The hyperparameters used for evaluation are detailed in Appendix B.

Evaluation Methodology The MTI Bench comprises 28 types of NLP tasks, yielding diverse outputs such as multiple-choice answers, numerical answers(fractional form), and extensive generative responses. Given this variety, directly applying verbalizers like LM-Eval-Harness (Gao et al., 2021) is impractical. Therefore, we prompted LLMs to return their outputs within an HTML tag (e.g., <task1>output<task1/>), which is then assessed via exact match (EM).

Hardware Specifications In Section 5.2, we examine the inference speed of four models: Tulu (7b, 13b, 30b, 65b) (Wang et al., 2023c). For observation, the hardware configuration for each model size is fixed. Specifically, the Tulu models with 7B and 13B parameters were tested using a single NVIDIA SXM4 with 80GB RAM. The 30B model utilized two of these NVIDIA SXM4 80GB GPU, while the largest, the 65B model, was evaluated using eight RTX A6000 with 48GB RAM each.

5 Experimental Results

In this section, we compare Single-Task Inference, Batching Prompting and Multi-Task Inference on the MTI Bench (Section 5.1), study the inference latency of each method (Section 5.2) and study the efficicacy of Multi-Task Inference on free-form generation (Section 5.3).

5.1 Main Results

We first evaluate Single-Task Inference, Batching Prompting, and Multi-Task Inference using the MTI Bench. In Table 5 we focus on the final accuracy of each model, only considering the cases where it correctly solves the entire combination of sub-tasks. Surprisingly, Multi-Task Inference consistently outperforms the other methods across various models. Notably, the performance gap between the inference strategies is larger in more powerful models. For instance, with the Llama-2-Chat-70B model, accuracy under Single-Task Inference and Batching Prompting is 8.7% and 7.9%, respectively, but it leaps to 16.0% using Multi-

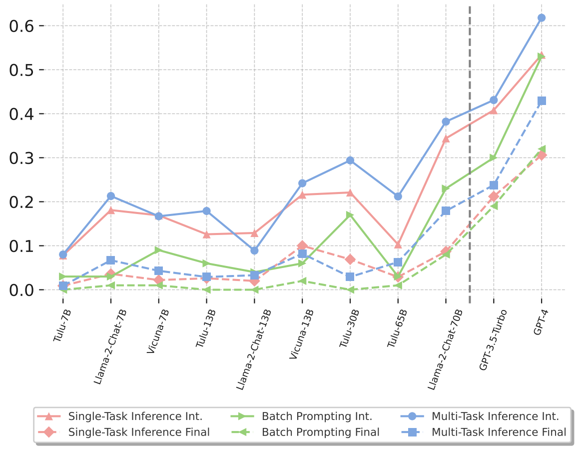

Figure 3: Comparative analysis of LLMs across SINGLE-TASK INFERENCE (Green), BATCH PROMPTING (Red), and MULTI-TASK INFERENCE (Blue). Solid lines represent the models' initial sub-task performance (i.e., **intermediate accuracy**), while dashed lines indicate their overall accuracy in completing the entire set of tasks (i.e., **final accuracy**). Models are listed in ascending order by parameter count, with proprietary models listed separately at the end.

TASK INFERENCE. A similar trend is observed in GPT-4, where accuracy escalates from 30.8% and 32.2% to 43.2%. In Figure 3, we observe a clear upward scaling trend, which demonstrates that more advanced models exhibit enhanced performance on the MTI BENCH, irrespective of the prompting methods employed. This trend suggests that the capability to concurrently handle multi-task instructions could be an *emergent* property (Wei et al., 2022a), associated with the increased scale of models.

The intermediate accuracy for each prompting method is illustrated in Figure 3. Notably, MULTI-TASK INFERENCE, depicted in blue, consistently surpasses alternative prompting methods in both initial and final performances. Furthermore, the efficacy of BATCH PROMPTING, depicted in green, improves as the model size increases, reaching its peak with GPT-4. Despite the improvement, however, a performance gap exists with the remaining inference methods. We conjecture that the performance margin may be tied to the operational nature of BATCH PROMPTING. It combines multiple tasks without regard to their inter-dependencies, potentially introducing unrelated contexts into a single prompt. This mixing of tasks can confuse the model, as it needs to navigate through irrelevant information multiple times to address the prompt accurately. This observation aligns with existing research that the performance of batching inference improves with model scale (Cheng et al., 2023) and that the presence of non-relevant context can adversely affect model performance (Shi et al., 2023).

Batch Size	N = 1		
Inference Type	SINGLE-TASK	BATCH PROMPTING	MULTI-TASK
TULU-7B	11.3 ± 5.6	5.1 ± 2.6	7.5 ± 5.2
TULU-13B	14.8 ± 6.0	6.3 ± 2.8	9.2 ± 5.4
TULU-30B	51.9 ± 61.9	46.2 ± 57.3	49.2 ± 42.2
TULU-65B	110.1 ± 54.1	52.6 ± 30.1	67.7 ± 39.6

Table 6: The inference latency in solving a multi-task instruction (with a batch size of 1) of the TULU models measured in seconds. This measurement is an average derived from 1,000 trials.

Batch Size	N = 4		
Inference Type	SINGLE-TASK	BATCH PROMPTING	MULTI-TASK
TULU-7B	15.4 ± 6.5	6.3 ± 2.8	11.3 ± 6.3
TULU-13B	19.4 ± 6.4	7.4 ± 3.1	13.0 ± 5.6
TULU-30B	93.3 ± 117.7	57.6 ± 65.1	63.2 ± 37.5
TULU-65B	156.9 ± 64.7	64.9 ± 33.8	96.7 ± 38.5

Table 7: The inference latency in solving a multi-task (with a batch size of 4) of the TULU models measured in seconds. This measurement is an average derived from 250 trials.

Finally, the MTI BENCH is divided into two subsets: MULTI-STEP and MULTI-PART. As seen in Table 5, models generally perform better in the MULTI-PART subset. This suggests that inter-task dependency in multi-task instructions is a significant factor that hinders LLM performance, and the ability to manage sequential task dependencies effectively is not uniformly developed across different models.

5.2 Inference Latency

Considering KV caching, intuitively, a model requiring fewer inference calls would be faster in terms of inference speed, assuming it generates an equal number of tokens. Empirically, in Tables 6 and 7, we observe a $1.46\times$ increase in speed using MULTI-TASK INFERENCE compared to SINGLE-TASK INFERENCE. This acceleration remains consistent as the batch size increases from 1 to 4.

Additionally, BATCH PROMPTING demonstrates a $2.1\times$ increase in speed compared to SINGLE-TASK INFERENCE, aligning with the findings in (Cheng et al., 2023). However, as highlighted in Section 5.1, employing BATCH PROMPTING for the MTI BENCH results in a marked decrease in performance, making MULTI-TASK INFERENCE the most viable option.

5.3 FREE-FORM GENERATION Subset

As mentioned in Section 4, LLMs are prompted to return their outputs within an HTML tag, which are parsed using regular expressions. During our evaluation, we notice that models often struggle to produce outputs in the correct format, potentially skewing their perceived performance. To address this issue, we introduce a new ablation subset called FREE-FORM GENERATION. This subset comprises 11 tasks, each divided into two sub-tasks, primarily focused on translation and summarization. Performance evaluation is conducted using the Rouge-L metric. Due to constraints in budget and time, this ablation is narrowed down to assess performance in the following methods: SINGLE-TASK INFERENCE and MULTI-TASK INFERENCE. Further details on the subset are provided in Appendix C.

Table 8 shows the result of our evaluation on the FREE-FORM GENERATION subset. We observe that smaller open-source models tend to perform better with SINGLE-TASK INFERENCE outperforming MULTI-TASK INFERENCE, with margins ranging from 0.02 to 0.15. However, this performance gap narrows for larger open-source models and proprietary models. Notably, for GPT-4, the difference in performance between the two methods is a mere 0.01, indicating that there is no significant difference in their effectiveness regardless of their output formatting.

We conjecture that the slight decrease in the performance of MULTI-TASK INFERENCE within the FREE-FORM GENERATION subset can be attributed to the weaker interdependence of the subtasks involved. For example, in task combinations such as translation and summarization, the information provided by the second instruction offers limited insights into solving the first task. This lack of inter-task informational clues may lead to a reduced level of synergy between the tasks, diminishing the benefit of MULTI-TASK INFERENCE in such scenarios.

In an effort to conduct a more comprehensive comparison between MULTI-TASK INFERENCE and SINGLE-TASK INFERENCE within free-form generation, we conducted further evaluations using the MT-BENCH (Zheng et al., 2023). A GPT-4 model, with the default pairwise comparison prompt from the original paper, was leveraged to judge and select the better response. The results, depicted in Figure 4, reveal that LLMs show a slightly improved performance under MULTI-TASK INFERENCE, with an average win rate of 58% across the prompts. Remarkably, LLAMA-2-CHAT-70B and GPT-4 under MULTI-TASK INFERENCE outperformed at 65.2% and 71.9% on the prompts, respectively. This shows that the benefit of MULTI-TASK INFERENCE persists beyond MTI BENCH and can be generalized to diverse use cases.

[2] https://pypi.org/project/rouge-score/

Models	FREE-FORM GENERATION			
	SINGLE-TASK		MULTI-TASK	
	1ST	FINAL	1ST	FINAL
TULU-7B	0.19	0.12	0.16	0.10
TULU-13B	0.17	0.21	0.15	0.12
TULU-30B	0.37	0.30	0.17	0.15
TULU-65B	0.39	0.11	0.19	0.15
LLAMA-2-CHAT-7B	0.14	0.10	0.11	0.08
LLAMA-2-CHAT-13B	0.15	0.13	0.07	0.16
LLAMA-2-CHAT-70B	0.35	0.26	0.27	0.24
VICUNA-7B	0.20	0.14	0.16	0.09
VICUNA-13B	0.22	0.18	0.23	0.16
GPT-3.5-TURBO	0.48	0.37	0.38	0.31
GPT-4	0.47	0.39	0.39	0.38

Table 8: Evaluation results of FREE-FORM GENERATION subset in SINGLE-TASK INFERENCE and MULTI-TASK INFERENCE. Evaluations are held in a one-shot setting. Note that four MCQA tasks are included in this subset as secondary tasks. Performance scores for both the generative and MCQA tasks are calculated using the Rouge-L metric.[2]

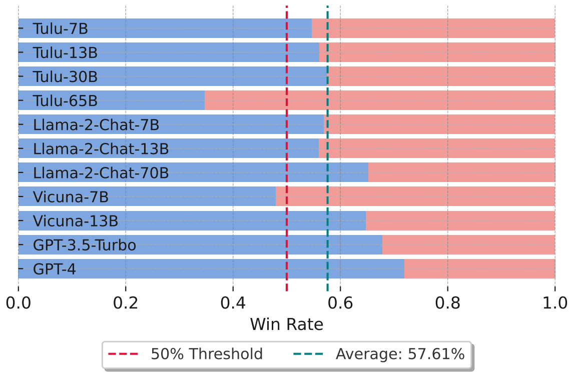

Figure 4: Win Rate Analysis. Blue bars represent MULTI-TASK INFERENCE wins, and red bars indicate SINGLE-TASK INFERENCE wins. The green line denotes the average MULTI-TASK INFERENCE win rate across all models.

6 Analysis of MULTI-TASK INFERENCE

In our previous section, although it is clear that MULTI-TASK INFERENCE guarantees speed-up (as explained in Section 5.2), it is rather unexpected and surprising that larger models show improved performance on the MTI BENCH (Table 5 and Figure 3) and the MT BENCH (Figure 4) compared to SINGLE-TASK INFERENCE. For a better understanding, we conduct an ablation experiment by inserting additional input components (Section 6.1) and conduct a human evaluation to categorize what would be the reason behind the performance improvement (Section 6.2).

	TULU-7B	TULU-13B	TULU-30B	GPT-3.5
		MULTI-STEP		
SINGLE-TASK INFERENCE	8.3	15.7	35	44.6
+ 2ND INSTRUCTION	8.5 (+0.2)	25.3 (+9.6)	36.2 (+1.2)	46.3 (+1.7)
+ 2ND CONTEXT	8.7 (+0.4)	20.5 (+4.8)	33.5 (-1.5)	46.3 (+1.7)
		MULTI-PART		
SINGLE-TASK INFERENCE	7.3	9.2	18.8	36.7
+ 2ND INSTRUCTION	6.5 (-0.8)	11.2 (+2.0)	18.0 (-0.8)	36.0 (-0.7)
+ 2ND CONTEXT	7.7 (+0.4)	11.4 (+2.2)	22.0 (+3.2)	38.7 (+2.0)

Table 9: Ablation experiment of excluding the 2ND INSTRUCTION and 2ND CONTEXT. The specified accuracy represents the models' performance on the first sub-task (denoted as **intermediate accuracy** in Figure 3). Note that SINGLE-TASK INFERENCE is excluding both the 2ND INSTRUCTION and 2ND CONTEXT compared to MULTI-TASK INFERENCE.

6.1 Ablation Experiment

A two-step instance within the MTI BENCH would consist of four input components when inferenced via MUTLI-TASK INFERENCE: (1) 1st Instruction, (2) 1st Context, (3) 2nd Instruction, (4) 2nd Context. On the other hand, when inferenced via SINGLE-TASK INFERENCE, only (1) 1st Instruction and (2) 1st Context would be provided as the input during the first inference call. Then on the second inference call, (3) the output of the first inference call, (4) 2nd Instruction, and (5) 2nd Context would be additionally provided.

In Table 9, we check the effect when adding the 2st Instruction and 2nd Context during SINGLE-TASK INFERENCE. Note that SINGLE-TASK INFERENCE is the same as excluding both the 2ND CONTEXT and 2ND INSTRUCTION. Interestingly, across different models and data subsets, we observe a consistent performance improvement when either the 2nd Instruction or the 2nd Context is provided as additional input when solving the 1st Instruction, indicating a sign of a look-ahead effect.

6.2 Qualitative Analysis

To analyze what kind of look-ahead effect might enable language models to show improved performance on the first instruction, we conduct a qualitative analysis by checking the 107 instances where GPT-4 correctly solves using MULTI-TASK INFERENCE but not with SINGLE-TASK INFERENCE. Interestingly, we discover the following four patterns that supplement the look-aheading behavior of LMs: (1) **No Outputs**: SINGLE-TASK INFERENCE provided no output, suggesting there were no viable answers. Conversely, MULTI-TASK INFERENCE, while acknowledging the implausibility of all answers, still opts to select one. (2) **Multiple Outputs**: SINGLE-TASK INFERENCE offered multiple answers, whereas the MULTI-TASK INFERENCE approach selected the most relevant one. (3) **Referencing**: MULTI-TASK INFERENCE leveraged information from a subsequent task to enhance its response to the initial task. (4) **Planning**: MULTI-TASK INFERENCE appeared to plan its solution before addressing the task.

	Observed Instances %
No Outputs	25%
Multiple Outputs	8%
Referencing	6%
Planning	3%

Table 10: Qualitative assessment results of GPT-4 outputs; The remaining 58% show no specific patterns.

Patterns 1 and 2 highlight the role of MULTI-TASK INFERENCE in providing a form of external feedback. The existence of subsequent tasks indicates whether an answer exists, thereby eliciting a response from the model. Conversely, Patterns 3 and 4 demonstrate that MULTI-TASK INFERENCE enables LLMs to utilize their full context window. This broader context usage, which extends beyond the immediate task, allows for more comprehensive problem-solving. The frequency of each pattern from our qualitative assessment is provided in Table 10. Sample instances of the observed patterns are provided as Figure 5.

7 Conclusion

In this work, we present the MTI BENCH, a comprehensive benchmark consisting of 5,000 instances spanning 25 diverse tasks, designed to assess the capability of LLMs in simultaneous multi-tasking. Our analysis within the benchmark compares MULTI-TASK INFERENCE, SINGLE-TASK INFERENCE and BATCH PROMPTING. The results indicate a superior performance by MULTI-TASK INFERENCE, despite reduced inference steps and a 1.46-fold increase in speed, demonstrating its efficiency in handling concurrent tasks.

8 Limitations

In this work, we try our best to offer a broad range of analyses, yet there are limitations that future studies should consider.

First, the MTI BENCH predominantly focuses on English, with the FREE FORM GENERATION ablation subset, adding French, and German. This linguistic range falls short of encompassing the wide diversity of different dialects and languages.

Second, the source dataset for MTI BENCH is largely oriented towards academic benchmarks. This focus might restrict its applicability in more general, user-oriented contexts. Future iterations should consider integrating more varied datasets to better mirror the multifaceted nature of everyday language use.

Third, another significant area concerns the automatic evaluation of model performance. Although our work employs a variety of methods such as model-based evaluation, exact matching, and Rouge-L, there is a need for additional studies on alignment with human preferences.

Fourth, the MTI BENCH only has a test set since the motivation was to test the MULTI-TASK INFERENCE capabilities of language models. Yet, it would be an interesting direction to see if the smaller models that underperformed in this work could improve their multi-processing capabilities by training on data instances with a similar format as the instances in the MTI BENCH.

Lastly, we only conducted our experiments in a one-shot setting. This was primarily because we observed that smaller models exhibit near zero accuracy when tested in a zero-shot setting and including more than two demonstrations resulted in too long input length. Yet, we acknowledge the importance of examining the impact of including additional demonstrations since models that support longer input lengths are gradually being introduced. We view this as a promising future research direction.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model apis. *arXiv preprint arXiv:2301.08721*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Pradeep Dasigi, Nelson F Liu, Ana Marasović, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv preprint arXiv:1908.05803*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Ari Holtzman, Peter West, and Luke Zettlemoyer. 2023. Generative models as a complex systems science: How can we make sense of large language model behavior? *arXiv preprint arXiv:2308.00189*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023a. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023b. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *arXiv preprint arXiv:1908.05852*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

David E Meyer and David E Kieras. 1997. A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms. *Psychological review*, 104(1):3.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data. *arXiv preprint arXiv:2301.11596*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Meeting of the Association for Computational Linguistics (ACL)*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences. *arXiv preprint arXiv:2106.00969*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

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

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. *arXiv preprint arXiv:2005.05763*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

A MTI Bench

Tables 13-37 provide a comprehensive overview of the 25 multi-tasks featured in the MTI BENCH. Each table includes the category, sub-tasks, and the original dataset of each multi-task. Furthermore, an example is provided to help a better understanding of the benchmark. Please note that for some examples, the context has been abbreviated for better readability.

B Inference Details

During our experiments, we use the hyperparameters as shown in Table 11.

Temperature	Top-p	Repetition Penalty	Max Output Length
0.7	1.0	1.0	2048

Table 11: Hyperparameters used for experiments in MTI BENCH.

C FREE FORM GENERATION Subset

Apart from MULTI-STEP and MULTI-PART subsets discussed in Section 3.1, we introduce a FREE-FORM GENERATION subset in our ablation studies detailed in Section 5.3. This subset follows the same creation process as the original MTI Bench, except for the hired annotators for quality assessment.

# of Tasks	12
# of Instances	2,400
Language	EN, FR, DE
List of Sub-Tasks	Translation, Summarization, MCQA
Source Dataset	FLORES-200 (Team et al., 2022) Belebele (Bandarkar et al., 2023) Wikilingua (Ladhak et al., 2020)

Table 12: Details for the FREE FORM GENERATION Subset

D Examples for Section 6.2

In this section we provide sample instances for the following patterns discussed at Section 6.2: No Outputs, Multiple Outputs, Referencing, and Planning. See Figure 5.

Task ID	001
Category	MULTI-STEP
Sub-Tasks	Answerability Classification - Extractive Question Answering
Source Dataset	QUOREF (Dasigi et al., 2019)
Example	Read the following passage, and follow the given steps, #1 Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task1>N<task1/>format. #2 Answer the question you have chosen in step #1. Return the answer in <task2>N<task2/>format. ### Context: Passage: Big Butte Creek drains approximately 245 square miles (635 km2) of southern Oregon. [. . .] ### List of Questions: (1) What watershed is split into two geographic regions? (2) What two entities was the foundation split into in october 2016? (3) What century was Europe split into two city states and kingdoms? (4) How many years was Nashua split into two cities? (5) Who likes to divide their projects into relevent time periods and geographic regions?

Table 13: Multi-Task 001 from the MTI BENCH.

Task ID	002
Category	MULTI-STEP
Sub-Tasks	Sentence Sorting - Bubble Sorting - Binary Question Answering
Source Dataset	SCRUPLES (Lourie et al., 2021)
Example	Read the following passage, and follow the given steps. #1 The list of sentences come after the context. Reorder them to its original order. Return the answer in <task1>(1)-(2)-(3)-(4)<task1/>format. #2 Use your answer for step#1 as a source sequence. Calculate the minimum number of swaps required to change the source sequence to (1)-(2)-(3)-(4). Swapping is only allowed between neighboring sequences. Return the answer in <task2>N<task2/>format. #3 Read the reordered text and solve the question. If the answer is "yes" multiply 2 to the answer for step#2. If the answer is "no" multiply 0.5. Return the answer in <task3>N<task3/>format. ### Context: **TL;DR My mother died this week, my girlfriend started a fight days later over my inability to talk about it, then she [. . .] ### List of sentences: (1) Or else am I justified in feeling this way? (2) I do try very hard to be self-critical, so if you all think I'm in the wrong here, need to just let this go or even apologize to her to prevent things from getting worse, then please tell me. (3) To me, she's being unnecessarily heated and petty on an awful week for me. (4) If she decides not to come, I feel like that might be the end of our relationship. ### Question: Is my girlfriend wrong in the post ?

Table 14: Multi-Task 002 from the MTI BENCH.

Task ID	003
Category	MULTI-STEP
Sub-Tasks	Answerability Classification - Multiple-Choice Question Answering
Source Dataset	COSMOS QA (Huang et al., 2019)
Example	Read the following passage, and follow the given steps, #1 Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task1>N<task1/>format. #2 Choose the correct answer for the question you have chosen in step #1. Return the answer in <task2>N<task2/>format. ### Context: Two cats (one is an itty bitty kitty that they bought a couple of days ago) , [. . .] ### List of Questions: (1) What makes it possible for humans to live on other planets? (2) What may be the reason there are so many people living in a small apartment? [. . .] ### Answer Choice: (1) The individual makes pictures all the time . (2) We all split the bills and makes it easier to live . [. . .]

Table 15: Multi-Task 003 from the MTI BENCH.

Task ID	004
Category	Multi-Part
Sub-Tasks	Answerability Classification - Answer & Question Matching
Source Dataset	DROP (Dua et al., 2019)
Example	Read the following passage, and follow the given steps, #1 Go through the provided list of questions and choose all that is answerable given the context. Return the answer in <task1>[N, N, ..]<task1/>format. #2 From the questions selected at task#1 choose the one that best suits the given answer. Return the answer in <task2>N<task2/>format. ### Context: Passage: Until 1998, Shearer was paid $30,000 per episode. During a pay dispute in 1998, [. . .] ### List of Questions: (1) Which year was the 400,000 salary per episode cut down by 100,000? (2) How many more dollars did voice actors receive in 2008 than they negotiated for in 2004? (3) How many years after taking the throne for himself and refusing to pay tribute did a military response begin? (4) How many years after receiving a raise did Shearer take a pay cut? (5) How many students does $16,000 a year pay for? ### Answer: 3.

Table 16: Multi-Task 004 from the MTI Bench.

Task ID	005
Category	Multi-Part
Sub-Tasks	Question & Context Matching - Wrong Candidate Ranking
Source Dataset	COSMOS QA (Huang et al., 2019)
Example	Read the following passage, and follow the given steps, #1 Read the following list of text and determine which one contains the answer to the question. Return the answer in <task1>N<task1/>format. #2 Read the list of wrong candidates provided determine which one serves as the best wrong answer for the question. Return the answer in <task2>N<task2/>format. ### Question: What does the narrator think about the video game they were playing ? ### List of Text: (1) The walk in was quite tiring actually plus the hot scorching sun. [. . .] (2) So basically the lecture was on when to know if the guy is a nutcase or not. [. . .] (3) I almost cried when I saw the mud in the arena , it was fucking insane! [. . .] ### Wrong Candidates: (1) She wants a PC. (2) Because it stopped running Firefox . (3) They lost it at school . (4) It could be a lot better . (5) They were taking a fitness test at the gym .

Table 17: Multi-Task 005 from the MTI Bench.

Task ID	006
Category	Multi-step
Sub-Tasks	Answerability Classification - Necessary Sentence Identification
Source Dataset	MULTI RC (Khashabi et al., 2018)
Example	Read the following passage, and follow the given steps, #1 Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task1>N<task1/>format. #2 Choose sentences from the context that is necessary to answer the question you have chosen in step #1. Return the answer in <task2>[N, N, ..]<task2/>format. ### Context: Sent 1: The film opens with Sunita , a medical student , and her friends working on a project about the human brain. Sent 2: She wants to investigate the curious case of Sanjay Singhania , a notable city businessman , who is reported to have anterograde amnesia. Sent 3: Her professor denies access to Sanjay 's records as it is currently under criminal investigation. [. . .] ### List of Questions: (1) can a person function with half a brain (2) Sunita is working on a project about the human brain and wants to interview which person with anterograde amnesia? (3) Beyonce did an interview with which magazine and was asked about feminism? (4) What is anterograde amnesia? (5) Why is the writer working on a project?

Table 18: Multi-Task 006 from the MTI Bench.

Task ID	007
Category	Multi-Part
Sub-Tasks	Sentence Sorting - Inappropriate Question Identification
Source Dataset	multi rc (Khashabi et al., 2018)
Example	Read the following passage, and follow the given steps, #1 The provided list of sentences come after the provided context, order the properly. Return the answer in <task1>(1)-(2)-(3)-(4)<task1/>format. #2 Choose one question that cannot be answered with the context. Return the answer in <task2>N<task2/>format. ### Context: Preservation and Conservation: In 1857 the Great Western Railway Company built a main line to Scotland, [. . .] ### List of Sentences: 1: In 1974 a total reorganization of local government throughout the UK did away with the old counties of Cumberland and Westmoreland and created the larger county of Cumbria. 2: While the Lake District encourages and welcomes visitors, its popularity can damage the landscape and tax local transportation services. [. . .] ### List of Questions: 1: What 1879 event caused a group of concerned individuals to form the Lake District Defense Association? 2: What organization was a precursor to the National Trust? [. . .]

Table 19: Multi-Task 007 from the MTI Bench.

Task ID	008
Category	Multi-Part
Sub-Tasks	Sentence Sorting - Answer & Question Matching
Source Dataset	ropes (Lin et al., 2019)
Example	Read the following passage, and follow the given steps, #1 The provided list of sentences come after the provided context, order them properly. Return the answer in <task1>(1)-(2)-(3)-(4)<task1/>format. #2 Choose one question that best suits the given passage and answer. Return the answer in <task2>N<task2/>format. ### Context: New species develop naturally through the process of natural selection. [. . .] ### List of Sentences: (a): Mike lives in a cold mid-western city, where there is not much predator prey interaction. (b): He also knew that darker coats are more suitable in cold environment with less predator prey interaction. [. . .] ### List of Questions: 1. Which squirrels would most likely reproduce in greater numbers, lighter or darker? 2. Would the color be darker or lighter at point B than at point A? [. . .] ### Answer: greater.

Table 20: Multi-Task 008 from the MTI Bench.

Task ID	009
Category	Multi-Step
Sub-Tasks	Necessary Sentence Identification - Sentence Sorting
Source Dataset	TIMETRAVEL (Qin et al., 2019)
Example	Read the following passage, and follow the given steps, #1 Choose one sentence that does not originally belong to the passage. Return the answer in <task1>N<task1/>format. #2 Reorganize the remaining sentences into its original order. Return the answer in <task2>(1)-(2)-(3)-(4)<task2/>format. ### List of Sentences: (1) My daughter jumped up and grabbed the blue one out of her hand (2) Nana chased her down, caught her, and tickled her until she laughed (3) She took off running down the hall while waving the sock in the air (4) She held up an orange sock and a blue one. (5) Nana came into the room with a puzzled look on her face. (6) She held up an orange shirt and a blue one.

Table 21: Multi-Task 009 from the MTI Bench.

Task ID	010
Category	Multi-Step
Sub-Tasks	Coherent Passage Detection - Sentence Sorting
Source Dataset	ABDUCTIVENLI (Bhagavatula et al., 2019)
Example	Read the following passage, and follow the given steps, #1 You will be given five group of sentences. Only one of them is a group of coherent sentences. The others include an injected sentence. Find the coherent passage. Return the answer in <task1>N<task1/>format. #2 Reorganize the passage you chose in step 1 into its original order. Return the answer in <task2>(1)-(2)-(3)-(4)<task2/>format. ### List of Sentences: 1. (1) Jackson now lives with the guilt of being a thief. (2) Mark kept the wallet. (3) Jackson stole a wallet at a party on Friday. 2. (1) The teacher also gave the lab partner detention for not doing anything. (2) The lab partner sat there like they knew everything. (3) The instructor announced the lab that we're going to perform.[. . .]

Table 22: Multi-Task 010 from the MTI Bench.

Task ID	011
Category	Multi-Step
Sub-Tasks	Question Classification - Multiple Choice Question Answering
Source Dataset	COMMONSENSEQA (Talmor et al., 2019)
Example	Read the following questions, and follow the given steps, #1 Choose one question that best suits a "CommonsenseQA" dataset. Return the answer in <task1>N<task1/>format. #2 Read the options and solve the question you chose at step#1. Return the answer in <task1>N<task1/>format. ### List of Questions: (1) What does the client think about the house? (2) Where would you put uncooked crab meat? (3) Why did the man buy dog food at the supermarket? (4) _, 52, earned about $94million in salary during his 16 seasons in the National Basketball Association. (5) Question: What is Hector Hammond's job?

Table 23: Multi-Task 011 from the MTI Bench.

Task ID	012
Category	Multi-Part
Sub-Tasks	Sentence Sorting - Answerability Classification - Extractive Question Answering
Source Dataset	SQUAD 1.1 (Rajpurkar et al., 2016)
Example	Read the following text, and follow the given steps, #1 Reorder the given sentences to its original order. Return the answer in <task1>(1)-(2)-(3)-(4)<task1/>format. #2 Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task2>N<task2/>format. #3 Solve the question you have chose from step#2. Extract the answer from the passage of step#1. Return the answer in <task3>N<task3/>format. ### List of Sentences: (1) The flowers tended to grow in a spiral pattern, to be bisexual (in plants, this means both male and female parts on the same flower), and to be dominated by the ovary (female part). (2) The most primitive flowers probably had a variable number of flower parts, often separate from (but in contact with) each other. [. . .] ### List of Questions: (1) Who'd tactic evolved? (2) When do they plant yams and millet? (3) What did some plant parts do as they evolved? (4) what became more mammal-like as they evolved? (5) What did some plant parts do when the flower had only male parts?

Table 24: Multi-Task 012 from the MTI Bench.

Task ID	013
Category	MULTI-PART
Sub-Tasks	Answer & Question Matching - Wrong Candidate Ranking
Source Dataset	PIQA (Bisk et al., 2020)
Example	Read the following text, and follow the given steps, #1 Choose the correct answer for the given question. Return the answer in <task1>N<task1/>format. #2 Choose the best incorrect answer for the given question. Return the answer in <task2>N<task2/>format. ### List of Answers: (1) Using a fork, stir the pecan mixture with the butter until evenly coated. Press pecan butter mixture into the bottom of your springform pan. (2) If the semolina mixture is too dry, you can add a few teaspoons of milk until it reaches the right consistency (3) Using a pie plate, stir the pecan mixture with the butter until evenly coated. Press pecan butter mixture into the bottom of your springform pan. (4) Heat up milk in the colander until it is 105 degrees, then add yeast and a pinch of sugar to the bowl of milk (5) Take some boiled milk in a small bowl and add the saffron strands to it and watch the saffron turn the milk yellow. ### Question: How do I add the pecan mixture in the pan when making creamy chocolate toffee torte?

Table 25: Multi-Task 013 from the MTI BENCH.

Task ID	014
Category	MULTI-STEP
Sub-Tasks	Classification - Arithmetic
Source Dataset	COM2SENSE (Singh et al., 2021)
Example	Read the following text, and follow the given steps, #1 Read through the following list of sentences and choose all sentences that are plausible and matches commonsense. Return the answer in <task1>[N, N,...]<task1/>format. #2 Count the number of inplausible sentences and express its ratio in fraction form. Return the answer in <task2>n/N<task2/>format. ### List of Sentences: (1) Natalie was embarrassed when her husband yelled at her in the store, so she told all her classmates about the experience. (2) It is better to have white wine with fish than red wine (3) Ricki was delighted to see that 2 customers came to her opening night. [. . .]

Table 26: Multi-Task 014 from the MTI BENCH.

Task ID	015
Category	MULTI-STEP
Sub-Tasks	Classification - Arithmetic - Arithmetic
Source Dataset	WINOWHY (Zhang et al., 2020)
Example	Read the following text, and follow the given steps, #1 Read through the following list of sentences and choose all sentences that are incorrect reasons for the given question. Return the answer in <task1>[N, N, ..]<task1/>format. #2 Count the number correct reasons and express its ratio in fraction form. Return the answer in <task2>n/N<task2/>format. #3 Solve the following equation: (ratio_of_correct_reason) add (ratio_of_wrong_reason) Write in decimal form. Return the answer in <task3>N<task3/>format. ### Question: Sentence: Carol believed that Rebecca suspected that she had stolen the watch. Question: Why does the 'she' refer to carol? ### List of Sentences: (1) Because If Rebecca regrets something of course she must of been the one that stole the watch. (2) Because Because rebecca wouldn't suspect herself in a crime, she would know. (3) Because Rebecca was known to have been in an abusive relationship with Carol. [. . .]

Table 27: Multi-Task 015 from the MTI BENCH.

Task ID	016
Category	MULTI-PART
Sub-Tasks	Classification - Classification - Multiple Choice Question Answering
Source Dataset	ARGUMENT FACET SIMILARITY CORPUS (Misra et al., 2016)
Example	Read the following text, and follow the given steps, #1 Read through the following list of texts. The topic of each text is one of the following: (1) death_penalty (2) gun_control (3) gay_marriage. Choose all text that suits the death_penalty topic. Return the answer in <task1>[N, N, ..]<task1/>format. #2 The type of each text is one of the following: (1) argument_similarity (2) argument_clarity. Out of the text you have chose in step#1 choose argument_clarity text. Return the answer in <task2>N<task2/>format. #3 Solve the question you have chose in step#2. Choose from: (1) Similar (2) Not Similar (3) Valid (4) Ivalid. Return the answer in <task3>N<task3/>format. List of Texts: (1) Sent1: Since heterosexuals are provided the means to have a happy marriage and homosexuals are not, homosexuals are not equal to heterosexuals. Sent2: Allowing straight marriage to provide for U.S. citizenship, while gays have no option (marriage or civil union). (2) Well, if that's a reason to ban homosexuals from marriage, then along the same line of thought, then any couple that is infertile or chooses not to have children should not be permitted to get married. (3) Sent1: The judge may or may not feel the death penaly is warranted. Sent2: Many people find some crimes heinous enough to warrent the death penalty.[. . .]

Table 28: Multi-Task 016 from the MTI BENCH.

Task ID	017
Category	MULTI-PART
Sub-Tasks	Sentence Sorting - Binary Question Answering
Source Dataset	MCSCRIPT (Ostermann et al., 2018)
Example	Read the following text, and follow the given steps, #1 The list of sentences come after the context. Reorder them to its original order. Return the answer in <task1>(1)-(2)-(3)-(4)<task1/>format. #2 Choose the best answer for the given question. Return the answer in <task2>N<task2/>format. ### Context: I find that cats are very good about reminding you when it is time for them to eat. They will meow and often stand by their bowl. [. . .] List of Sentences: (1) So the first thing I do is head to the kitchen to see if there is an open can of her food in the refrigerator. (2) I am careful to measure her food so that she gets just a quarter cup of wet and a quarter cup of dry because I don't want her to be overweight. (3) Next I 'll go to my pantry and pull out a bag of her favorite dry food and mix a little of each into her food bowl. (4) Then I 'll take the time to make sure she has plenty of water before I set her dish on the floor for her to begin eating. ### Question: What is taken from the kitchen cupboard? Options: 1: measuring cup 2: Bag of cat food.

Table 29: Multi-Task 017 from the MTI BENCH.

Task ID	018
Category	MULTI-STEP
Sub-Tasks	Necessary Sentence Identification - Sentence Sorting - Extractive Question Answering
Source Dataset	DUORC (Saha et al., 2018)
Example	Read the following passage, and follow the given steps. #1: The list of sentences come after the context. Choose one that does not original belong to the context. Return the answer in <task1>N<task1/>format. #2: Reorder the remaining into its original order. Return the answer in <task2>(1)-(2)-(3)-(4)<task2/>format. #3: Answer the given question. Return the answer in <task3>N<task3/>format. ### Context Deepak (Shashi Kapoor) is on trial for the murder of his wealthy wife Vimla, but is acquitted and set free. [. . .] ### List of Sentences (1) Soon, Sapna learns that Gopal's real name is Deepak, who was previously accused of murdering his first wife. (2) Gopal's ever changing behavior throws everyone into suspicion and Sapna fears she will be his next victim. (3) Is Gopal innocent or Guilty? [. . .] ### Question Who believes that Gopal is annoying and is stalking her?

Table 30: Multi-Task 018 from the MTI BENCH.

Task ID	019
Category	Multi-part
Sub-Tasks	Natural Language Inference - Natural Language Inference
Source Dataset	SNLI (Bowman et al., 2015b)
Example	Read the following text, and follow the given steps, #1 Determine the relationship. between sentences 1&2. Choose from: (1) Entailment (2) Contradiction (3) Neutral. Return the answer in <task1>N<task1/>format. #2 Choose between the given list of sentences that replaces sentence 2 and make a entailment relationship with sentence 1. Return the answer in <task2>N<task2/>format. ### Sentence 1: An older man, dressed in red, yellow, and black, is standing outside waving a large flag and a long horn. ### Sentence 2: An older man is standing outside waving to a car driving past. ### List of Sentences: (A) An older man is proudly waving a large American flag. (B) There is a man outdoors waving a flag.

Table 31: Multi-Task 019 from the MTI Bench.

Task ID	020
Category	Multi-part
Sub-Tasks	Classification - Natural Language Inference
Source Dataset	MNLI (Williams et al., 2018)
Example	Read the following text, and follow the given steps, #1 Classify the given statements to one of the following categories : 1. FACE-TO-FACE, 2. GOVERNMENT, 3. LETTERS, 4. 9/11, 5. SLATE, 6. TELEPHONE, 7. TRAVEL, 8. VERBATIM, 9. OUP, 10. FICTION. Choose all that fits in category 5. Return the answer in <task1>[N, N, ..]<task1/>format. #2 Choose a sentence that is in an entailment relationship with the statement you chose in step#1. If their are two or more answer for step#1 use the first one. Return the answer in <task2>N<task2/>format. List of Statements: (1) yes but yes and i kind of have always pooh-poohed military educations but i think that for this kid [. . .] (2) He was pro-German, as he would have been pro-Boer. (3) Historian Thomas Reeves believes that, despite the media's reluctance to look into Kennedy's private life, if he had lived to have a second [. . .] List of Sentences: 1. This kid is not very well behaved or smart. 2. I generally don't like the idea of military educations. 3. I fully support military educations for kids.

Table 32: Multi-Task 020 from the MTI Bench.

Task ID	021
Category	Multi-part
Sub-Tasks	Algebra - Differentiation
Source Dataset	Super Natural Instructions - task 090 (Wang et al., 2022)
Example	Read the following passage, and follow the given steps. #1 Solve the given equation: $3 + 8x^1 + 6x^2$, x=10. Return the answer in <task1>N<task1/>format. #2 Differentiate the equation from step#1 Solve the equation. Return the answer in <task2>N<task2/>format.

Table 33: Multi-Task 021 from the MTI Bench.

Task ID	022
Category	Multi-Step
Sub-Tasks	Prime Classification - Arithmetic
Source Dataset	Super Natural Instructions - task 092 (Wang et al., 2022)
Example	Read the following passage, and follow the given steps. #1 Choose all prime numbers: (1) 99028 (2) 41549 (3) 51481 (4) 94135. Return the answer in <task1>[N, N, ...]<task1/> format. #2 Sum your choices at step#1.Return the answer in <task2>N<task2/> format.

Table 34: Multi-Task 022 from the MTI Bench.

Task ID	023
Category	Multi-Step
Sub-Tasks	Classification - Arithmetic
Source Dataset	MATH (Hendrycks et al., 2021)
Example	Read the following passage, and follow the given steps. #1 Read through the given questions. Each question fall into one of the following categories. Choose a question of measurement category. Return the answer in <task1>N<task1/>format. #2 Solve the question you have chose at step#1. Return the answer in <task2>N<task2/>format. ### List of Questions: (1) What is prob of picking 1 h and 2 p when three letters picked without replacement from {h: 1, e: 3, p: 2, n: 6, q: 1}? (2) Let p = 182843/22 + -8316. Calculate the common denominator of 70/32 - (1 + -1) and p. (3) How many milliseconds are there in 38.5396 microseconds? (4) Let y(a) = -a + 5. Let m be y(3). Solve f + 16 = -0*f - 4*c, -3*c - 12 = -m*f for f. (5) Calculate (3/(-6))/(33/(-44)).

Table 35: Multi-Task 023 from the MTI Bench.

Task ID	024
Category	Multi-Step
Sub-Tasks	Classification - Multiple Choice Question Answering
Source Dataset	MMLU (Hendrycks et al., 2020)
Example	Read the following passage, and follow the given steps. #1 Read through the given questions. Choose one question that is high school level. Return the answer in <task1>N<task1/>format. #2 Solve the question you have chose at step#1. Return the answer in <task2>N<task2/>format. List of Questions: (1) A discrete graph is complete if there is an edge connecting any pair of vertices. How many edges does a complete graph with 10 vertices have? (A)10 (B)20 (C)25 (D)45 (2) When n = 11, what is the value of 10 – (n + 6)? (A)–7 (B)5 (C)7 (D)27 (3) Find the area of the first quadrant region bounded by y = x^2, y = cos(x), and the y-axis. (A)0.292 (B)0.508 (C)0.547 (D)0.667

Table 36: Multi-Task 024 from the MTI Bench.

Task ID	025
Category	Multi-part
Sub-Tasks	Mask Infilling - Judicial Decision
Source Dataset	CASEHOLD (Zheng et al., 2021)
Example	Read the following passage, and follow the given steps. #1 Read through the given text. Choose one phrase that best suits the blank([MASK]) in the text. Return the answer in <task1>N<task1/>format. #2 Given the text choose the correct judicial decision. Return the answer in <task2>N<task2/>format. ### Context: Denial of his motions in Leo, Mr. Leo filed his second suit in the Kansas federal district court (No. 10-2495-JTM), [. . .] In a memorandum and order dated March 22, 2011, 2011 WL [MASK] [MASK] [MASK] [MASK] concluded that [. . .] ### List of Phrase: (1) contemporaneous objections. See, e.g., (2) the reality and the (3) substantially certain to result (4) 1097759, the district court (5) the City bears the ### List of Judicial Decisions (1) holding that although pro se briefs are liberally construed even pro se litigants must brief arguments in order to preserve them (2) holding that pro se pleadings are to be liberally construed [. . .]

Table 37: Multi-Task 025 from the MTI Bench.

No Outputs

Instruction #1
Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task1>N<task1/> format.

Context :
Brady 's mouth stiffened into a grim line , and he rubbed his mouth with his hand , not knowing what to do or [...]

Questions:
(1) if he can help, what will be awarded to him?
(2) What is likely to happen to Brady after acting this way in front of everyone ? [...]

Single-Task Inference:

Evaluation of the questions:
(1) The context mentions the place Jin Wu and his wife escaped to.
(2) The context mentions the person Jin Wu holds [...]
Based on the above, the questions that can be answered using the context are:
<task1>1<task1/>, <task1>2<task1/>, and <task1>3<task1/>.

Mutli-Task Inference:

From the context, it can be observed that the information provided is primarily about Brady's emotional state, [...]

None of the answers perfectly fit the context and the chosen question, but the closest one is: <task2>2<task2/>.

Multiple Outputs

Instruction #1
Go through the provided list of questions and choose the one that is answerable given the context. Return the answer in <task1>N<task1/> format.

Context :
Sent 1: In Japan of the Tokugawa Ieyasu period , a young ninja named Jin Wu wantonly kills samurai and other [. . .]
Sent 2: When they start hunting him down , Jin Wu and his wife Akane sail to [. . .]

Questions:
(1) Where did Jin Wu and his Wife escape to? [. . .]

Single-Task Inference:

Evaluation of the questions:
(1) The context mentions the place Jin Wu and his wife escaped to.
(2) The context mentions the person Jin Wu holds [...]
Based on the above, the questions that can be answered using the context are:
<task1>1<task1/>, <task1>2<task1/>, and <task1>3<task1/>.

Mutli-Task Inference:

Evaluation of the questions:
(1) The context mentions where Jin Wu and his wife escaped to.
(2) The context also mentions [. . .]

Based on the above, all the questions are answerable using the context. But for this task, we will choose only one: <task1>2<task1/>

Referencing

Instruction #1
Choose CommonsenseQA question from the provided options. Return the answer in <task1>N<task1/> format.

###List of Questions
[...]
(5) Where are you when you're about to use your plane ticket?
###Options:
(A)pocket (B)terrorists hands (C)airport (D)sea ship (E)briefcase

Single-Task Inference:

Question Analysis:
[...]
Question 4 is commonsenseQA question requiring knowledge on a specific product to solve. [...]
The answer is <task1>4<task1/>

Mutli-Task Inference:

Question Analysis:
[...]
(5) Where are you when you're about to use your plane ticket? - This is a practical, common sense-based question. **The options like airport seem relevant here.**

The answer is <task1>5<task1/>.

Planning

Instruction #1
Choose the correct instruction for solving the given task. Return the answer in <task1>N<task1/> format.

List of Instructions
1) Poke holes in the foil lid with a toothpick.
(2) Take a few sips of pickle juice.

'###Task
Reduce heartburn pain.

Single-Task Inference:

The correct instruction is "Take a few sips of pickle juice" (Instruction 2). While this might seem unconventional, pickle juice has been suggested as a home remedy for heartburn. It contains vinegar, which is thought to help by neutralizing stomach acid, thus [. . .]

<task1>2<task1/>

Mutli-Task Inference:

1. **Step 1 - Pinpoint the Helpful Instruction:** Identify the step that most effectively reduces heartburn pain.

2. **Step 2 - Identify the Harmful Instruction**: From the remaining steps, find the one that could potentially aggravate heartburn, and then sum this number with the one from Step 1.

[. . .]

Figure 5: Examples of **No Outputs**, **Multiple Outputs**,**Referencing** and **Planning** used by GPT-4 during the MULTI-TASK INFERENCE.