Inflection-1

Inflection AI
techmemo@inflection.ai

June 22, 2023

Model	MMLU (5)	HellaSwag	PIQA	BoolQ	Natural QA	GSM8K (5)
LLaMA 65B	63.4	84.2	82.8	85.3	23.8	50.9
Chinchilla	67.5	80.8	81.8	83.7	16.6	-
PaLM 540B	69.3	83.4	82.3	88.0	21.2	56.6
GPT-3.5	70.0	-	-	-	-	57.1
Inflection-1	**72.7**	**84.3**	**84.2**	**89.7**	**29.8**	**62.9**

Table 1: **Comparison to models in the Inflection-1 compute class.** We show a comparison to models that are in the same compute class as Inflection-1, which we consider to be models trained with at most the training FLOPs of PaLM 540B. We find that Inflection-1 outperforms these models across a wide range of benchmarks. All evaluations are 0-shot unless shown in parentheses.

Introduction

Large language models (LLMs) based on the Transformer architecture have been shown to possess a range of advanced capabilities in language generation and understanding. These capabilities have paved the way for deployment of LLMs in products like OpenAI's ChatGPT and Google's Bard. At Inflection AI, our mission is to create personal AIs for everyone, and in May 2023 we released Pi (pi.ai) – an LLM-based personal AI which is designed to be empathetic, useful, and safe. In this work we introduce the foundation model powering Pi, dubbed Inflection-1, and evaluate its performance characteristics across a variety of benchmarks.

We find that Inflection-1 outperforms well-known models like GPT-3.5, LLaMA, PaLM 540B, and Chinchilla on a large number of benchmarks. Inflection-1 is the best performing model in its compute class, behind only PaLM-2 (L) and GPT-4 overall. In the following sections we describe these results in detail.

It is worth noting that a foundation LLM typically undergoes a complex adaptation process, such as alignment with human preferences and the safety policy, before it can be deployed in a user-facing product. Pi is no exception, which means that some of Inflection-1's capabilities are enhanced in Pi, while others are suppressed. The evaluation that we present in this memo covers Inflection-1, rather than Pi, and our choice of benchmarks is primarily focused on measuring its knowledge and reasoning capabilities. Safety is imbued in Pi at a later adaptation stage, which we will describe in a separate memo.

Compute Classes

To offer a fair comparison amongst models of varying sizes and training methods, we segmented foundation models into those pretrained using at most the FLOPs of Google's PaLM-540B (approximately 10x GPT-3) and those which used more. Models in the former category are usually faster to serve and can be deployed more widely, and include well-known models like LLaMA and Chinchilla. Models in the latter category tend to have the highest performance. When pretraining FLOPs are not reported publicly, we make a reasonable guess (for example, assigning GPT-3.5 to the former category and GPT-4 to the latter).

Inflection-1 was trained on a large dataset using thousands of NVIDIA H100 GPUs, and is a model within the first compute class. As such we focus our evaluations on this setting, though we include comparisons to models in the second compute class when benchmarks are available. The Inflection-1 architecture, dataset, and training procedure are proprietary, and we omit their details in this memo.

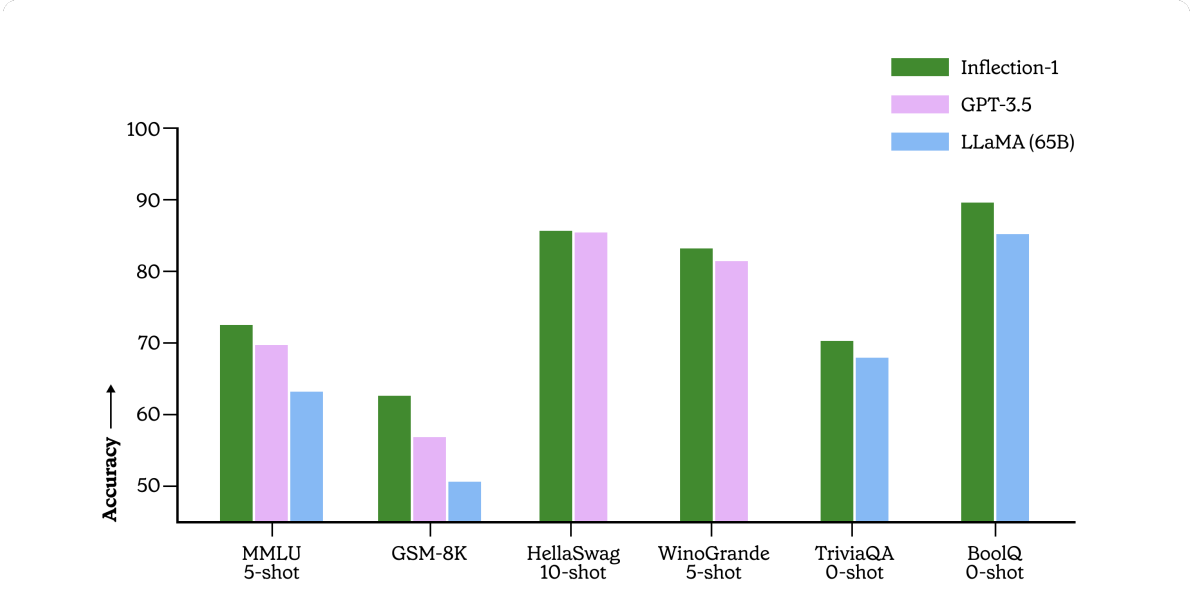

Figure 1: **Overview of Inflection-1's performance relative to LLaMA and GPT-3.5, two commonly deployed LLMs within the same compute class as Inflection-1.** In the text below, we compare to many more models, including GPT-4, PaLM 540B, PaLM 2, and Chinchilla.

Results

Below we present Inflection-1's performance on a set of commonly used benchmarks including common sense tasks, question answering, knowledge intensive tasks, reading comprehension, and code generation.

For Inflection-1, we report results without instruction tuning or RLHF. Other work has shown that these methods can be used to improve performance on specific benchmarks (1; 2), though we focus this work on our pre-trained model without any fine-tuning. When referencing GPT-4 and GPT-3.5, we show values from (3).

Multitask Language Understanding

In Table 2 we show results on Massive Multitask Language Understanding (**MMLU**) (4), a diverse collection of 57 tasks comprising high school, college, and professional level exams. We show 5-shot results as this allows us to compare to a wide range of models.

Model	Average	Humanities	STEM	Social Sciences	Other
GPT-4	86.4	-	-	-	-
PaLM 2-L	78.3	-	-	-	-
GPT-3.5	70.0	-	-	-	-
PaLM (540B)	69.3	77.0	55.6	81.0	69.6
Chinchilla (70B)	67.5	63.6	54.9	79.3	73.9
LLaMA (65B)	63.4	61.8	51.7	72.9	67.4
Inflection-1	72.7	79.2	61.7	82.6	74.1

Table 2: **5-shot results on MMLU comparing Inflection-1 to a wide range of models.** When applicable we show average results for the different categories laid out in (4).

Closed Book Question Answering

We show results on **TriviaQA** (5) along with **Natural Questions** (6) below. We follow the same evaluation format as used in (7). For TriviaQA, we use the dev set of the unfiltered set which is available online and we are able to compare to Chinchilla and LLaMA. We also report our 1-shot evaluation on the same split as reported in (2) to compare to PaLM-2 L.

Model	TriviaQA 0-shot	TriviaQA 1-shot	NaturalQA 0-shot	NaturalQA 1-shot
PaLM 2-L	-	-	-	37.5
Chinchilla (70B)	55.4	-	16.6	-
PaLM (540B)	-	-	21.2	29.3
LLaMA (65B)	68.2	71.6	23.8	31.0
Inflection-1	70.3	73.6	29.8	35.9

Table 3: **0-shot and 1-shot results on closed-book question answering tasks.** Results and formatting taken from (2; 7). For TriviaQA we show results for the same split as in (7), which is a different split than reported by PaLM-2 L. Using the same split as PaLM-2 L, Inflection-1 achieves 85.0% accuracy 1-shot compared to PaLM-2 L which achieves 86.1%.

Common Sense Benchmarks

We evaluate Inflection-1 on a variety of common sense benchmarks and compare against other large language models. We include **HellaSwag** (8), **PIQA** (9), **WinoGrande** (10), and **BoolQ** (11).

In Tables 4 and 5, we show 0-shot and k-shot results, respectively.

Model	HellaSwag	PIQA	BoolQ
PaLM (540B)	83.4	82.3	88.0
Chinchilla (70B)	80.8	81.8	83.7
LLaMA (65B)	84.2	82.8	85.3
Inflection-1	84.3	84.2	89.7

Table 4: **0-shot results on common sense benchmarks.** We compare to PaLM 540B (1), Chinchilla (12), and LLaMA (7).

Model	BoolQ 1-shot	HellaSwag 1-shot	HellaSwag 10-shot	WinoGrande 5-shot
GPT-4	-	-	95.3	87.5
PaLM 2-L	90.9	86.8	-	90.9
GPT-3.5	-	-	85.5	81.6
PaLM (540B)	88.7	83.6	-	85.1
PaLM 2-M	88.6	84.0	-	-
Inflection-1	88.9 (89.7 0-shot)	84.4	85.8	83.3

Table 5: **k-shot results on common sense benchmarks with comparison to GPT-4 and PaLM 2.** We note that for BoolQ we get worse results 1-shot than 0-shot.

BIG-Bench Hard

We show BIG-Bench Hard (13) with Chain-of-Thought prompting results in Table 6 using the setup described in (14).

Model	Results with CoT
PaLM 2-L	78.1
PaLM (540B)	65.2
Inflection-1	69.9

Table 6: **Results on BIG-Bench hard with Chain of Thought prompting.** We follow the same evaluation protocol as in (14) using the prompt from `https://github.com/suzgunmirac/BIG-Bench-Hard`.

Reading Comprehension

In Table 7, we show results on **RACE** (17) and **LAMBADA** (18). We evaluate RACE in the same format as used in the evaluation of Chinchilla and Gopher.

Mathematical Reasoning

We evaluate Inflection-1 on **GSM8K** (19), which contains grade school math word problems, and **MATH** (4), a dataset of high school competition problems divided in 7 subject areas. For MATH,

Model	LAMBADA 0-shot	LAMBADA 1-shot	RACE-m	RACE-h
PaLM-2 L	-	86.9	-	-
PaLM-2 M	-	83.7	-	-
PaLM (540B)	77.9	81.8	-	-
Chinchilla (70B)	77.4	-	86.8	82.3
Inflection-1	78.5	83.3	93.3	88.9

Table 7: **We show results on the reading comprehension benchmark RACE along with LAMBADA.** We include both 0 and 1 shot results on LAMBADA to allow us to compare to PaLM 2 and to other models. RACE is evaluated as in (15; 12). We evaluate 1-shot LAMBADA using the format outlined in (16).

we follow (20) and use the same 4-shot chain-of-thought prompt to generate answers. Generated answers are compared to gold references with the `SymPy` library (21) to take into account equivalent results represented differently. For GSM8K, we use the 8-shot chain-of-thought prompt from (22). In Table 8, we report results of Inflection-1 compared to other language models, including GPT-4, which was trained on a fraction of the training set of GSM8K and MATH.

Model	GSM8K	MATH
GPT-4	92.0	-
PaLM 2-L	80.7	34.3
GPT-3.5	57.1	-
PaLM (540B)	56.6	8.8
LLaMA (65B)	50.9	10.6
Inflection-1	62.9	16.7

Table 8: **Results on mathematical reasoning datasets.** We report results on GSM8K using the 8-shot prompt from (22) and on MATH using the 4-shot prompt from (20). On MATH the generated answer is compared to the gold reference with `SymPy`.

Code Generation

We evaluate the ability of Inflection-1 to generate code from a natural language description on **HumanEval** (23) and **MBPP** (24). The natural language description is presented as a Python docstring in HumanEval, while in MBPP it is a natural language instruction containing test cases. We evaluate Inflection-1 in a 0-shot setting on HumanEval, and use a 3-shot prompt similar to (24) for MBPP. A generated function is counted as correct if it passes the pre-defined tests. In Table 9, we report pass@1 scores of Inflection-1 compared to other models which are not trained or fine-tuned specifically for code. We find code generation to be the only benchmark where our model underperforms GPT-3.5. As our products do not require advanced code generation, we did not perform any work to specifically improve coding capabilities. It is possible to improve the ability of language models to generate code by fine-tuning on code-specific data (1; 23; 25; 26).

Model	HumanEval @1	MBPP @1
GPT-4	67.0	-
GPT-3.5	48.1	-
PaLM (540B)	26.2	36.8
LLaMA (65B)	23.7	37.7
Inflection-1	35.4	43.8

Table 9: **Results on code generation tasks.** For MBPP we report pass@1 scores using the 3-shot prompt from (24). For HumanEval we show results in the 0-shot setting.

Conclusion

In this technical memo we have shown that Inflection-1 – a foundation LLM developed by Inflection AI – outperforms well-known models such as GPT-3.5, LLaMA, PaLM 540B, and Chinchilla on a large number of benchmarks. Inflection-1 is the best performing model in its compute class, behind only PaLM-2 (L) and GPT-4 overall. A variant of Inflection-1 that has undergone alignment with human preferences and incorporates a safety policy is deployed in the personal AI product Pi. We believe that further advancements in AI capabilities and safety will usher in new products and user experiences, and we are committed to continuously improving our AI models.

References

[1] A. Chowdhery et al. Palm: Scaling language modeling with pathways, 2022.

[2] R. Anil et al. Palm 2 technical report, 2023.

[3] OpenAI. Gpt-4 technical report, 2023.

[4] D. Hendrycks et al. Measuring massive multitask language understanding, 2021.

[5] M. Joshi et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, July.

[6] T. Kwiatkowski et al. Natural questions: a benchmark for question answering research. Transactions of the Association of Computational Linguistics, 2019.

[7] H. Touvron et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[8] R. Zellers et al. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019.

[9] Y. Bisk et al. Piqa: Reasoning about physical commonsense in natural language, 2019.

[10] K. Sakaguchi et al. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 64(9):99–106, 2021.

[11] C. Clark et al. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019.

[12] J. Hoffmann et al. Training compute-optimal large language models, 2022.

[13] A. Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023.

[14] M. Suzgun et al. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022.

[15] J. W. Rae et al. Scaling language models: Methods, analysis & insights from training gopher, 2022.

[16] T. B. Brown et al. Language models are few-shot learners, 2020.

[17] G. Lai et al. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683, 2017.

[18] D. Paperno et al. The lambada dataset: Word prediction requiring a broad discourse context, 2016.

[19] K. Cobbe et al. Training verifiers to solve math word problems, 2021.

[20] A. Lewkowycz et al. Solving quantitative reasoning problems with language models, 2022.

[21] A. Meurer et al. Sympy: symbolic computing in python. PeerJ Computer Science, 3:e103, January 2017.

[22] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[23] M. Chen et al. Evaluating large language models trained on code, 2021.

[24] J. Austin et al. Program synthesis with large language models, 2021.

[25] E. Nijkamp et al. Codegen: An open large language model for code with multi-turn program synthesis, 2023.

[26] R. Li et al. Starcoder: may the source be with you!, 2023.