WRITER

Becoming self-instructed - The key to building high-quality models

Waseem AlShikh

CTO & Co-founder

L'ORÉAL Vanguard Johnson&Johnson INTUIT HubSpot

NVIDIA GTC

WRITER

The generative AI platform for enterprises

Headquarters
San Francisco

Founded
2020

Metrics
100K+ users
200+ customers

Investors
ICONIQ, Insight, WndrCo, Balderton, Google

CUSTOMERS

How we started

WRITER

We used the best training data

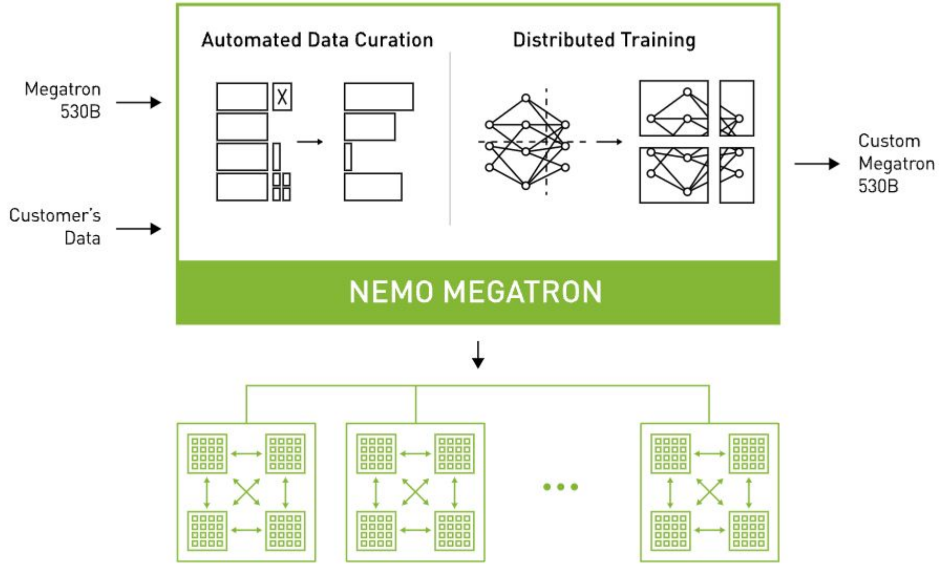
Lots of GPUs!

A few other tricks such as multiquery attention

Our results was much lower than expected.
Our instruct model was worse than plain vanilla LLM...

NeMo framework - NEMO MEGATRON

WRITER

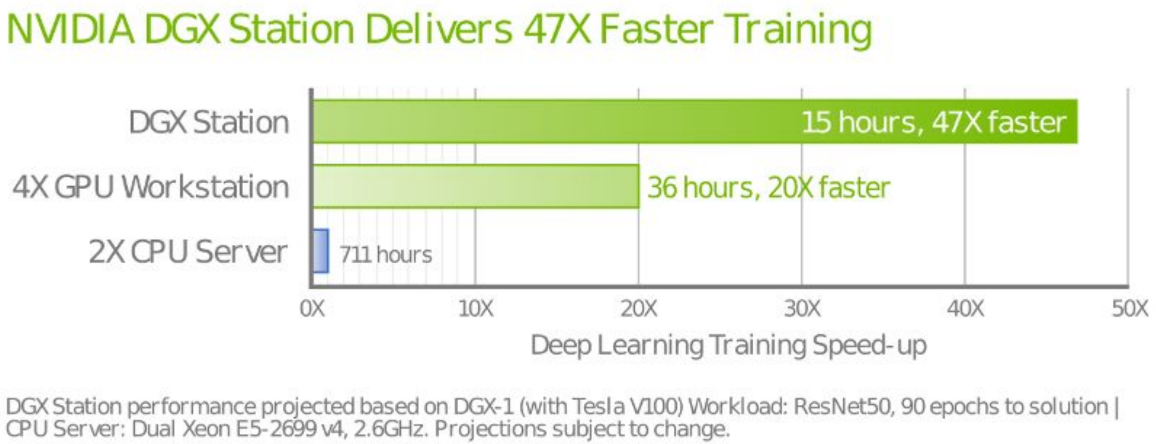


We've harnessed the power of Nemo-Megatron for training our large language models.

Its built-in distributed training capabilities, coupled with prebuilt libraries, make LLM training and deployment seamless.

NVIDIA DGX

WRITER



Our deep learning projects leverage NVIDIA GPUs for both model training and inference.

DGX Station delivers 3X the training performance of today's fastest workstations.

NVIDIA GPUs and associated tools are at the heart of our success in training and fine-tuning large language models.

They have not only expedited our projects but have also opened doors to previously unattainable performance levels.

https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/nlp/megatron.html

Data and methods

Becoming self-instructed data and training

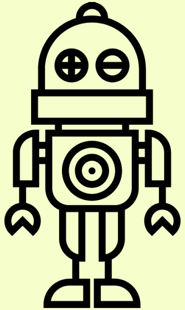
Aligning LLMs with user intent

The tale of two models

WRITER

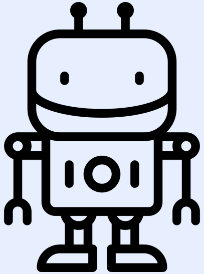
Vanilla model (non instruct)

Also known as a base model



“Let me continue your question”

Instruct model

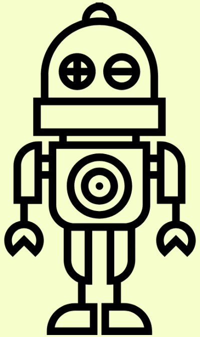


“Let me answer your question”

Types of common vanilla models

WRITER

Next word prediction



“Let me continue your question”

WRITER

Palmyra-base
Palmyra-large

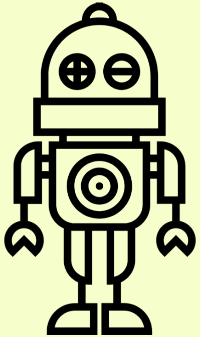
GPT2
GPT3

RedPajama
LLaMA
OpenLLama
OTB
MTB
…

Vanilla model predicting the next word

WRITER

Next word prediction



"Let me continue your question"

Non instruct

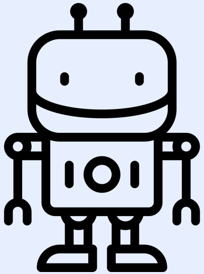
QUESTION:
What is the capital of France?

ANSWER:
What is the capital of Germany?

Types of common instruct models

WRITER

Follow instruction



“Let me answer your question”

WRITER

Palmyra-instruct
Palmyra-X
Camel

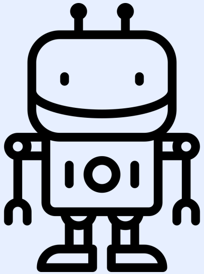
ChatGPT
GPT4
GPT3.5-turbo
InstructGPT

Alpaca
Vicuna
OpenChat
Orca
…

Instruct model answers your question

WRITER

Follow instruction



“Let me answer your question”

Instruct

QUESTION:

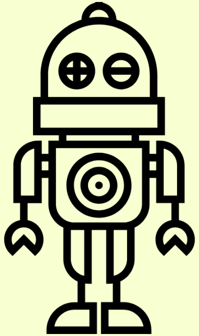
What is the capital of France?

ANSWER:

The capital of France is Paris.

Two tasks

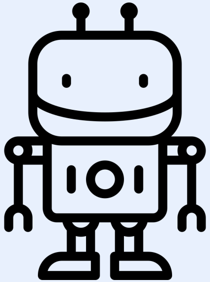
Next word prediction



“Let me continue your question”

Instruct tuning

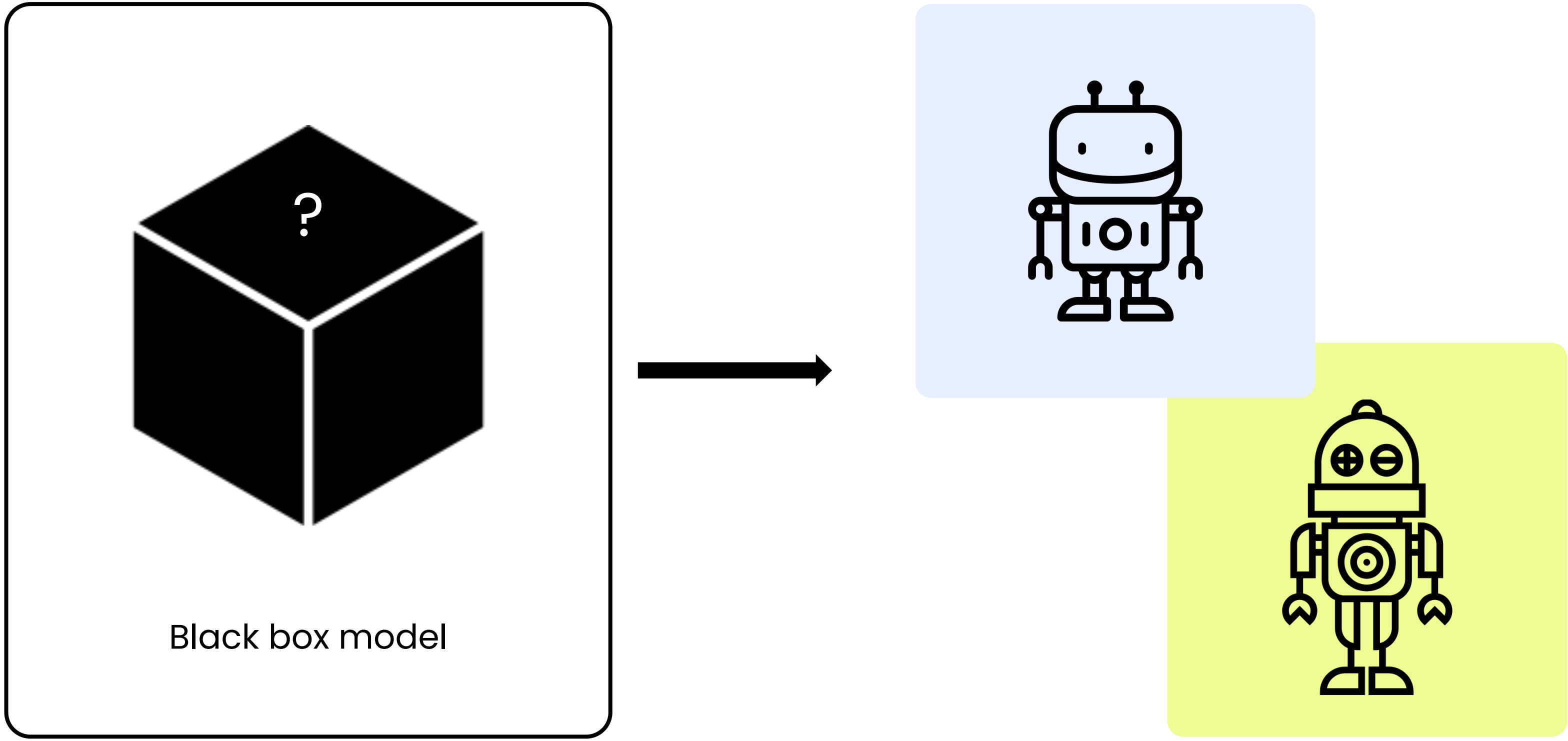
Follow instruction



“Let me answer your question”

Is the black box model vanilla or instruct?

WRITER

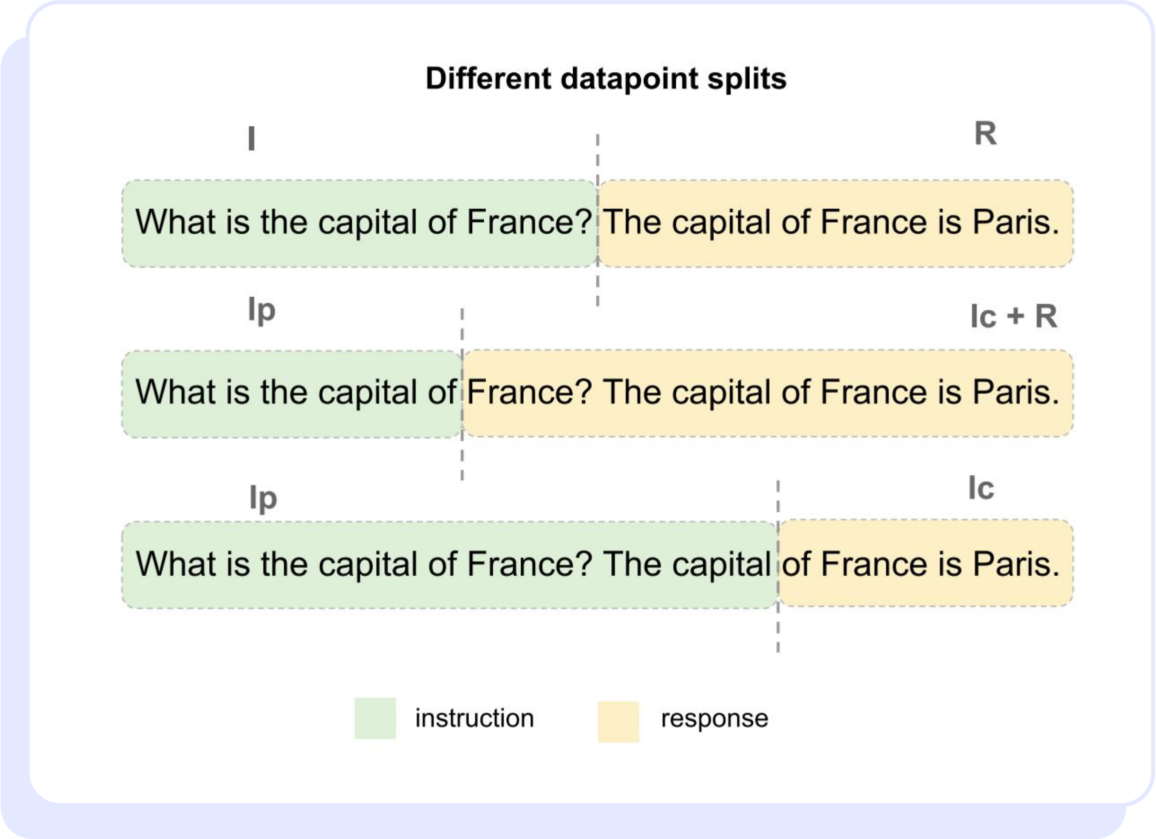


Introducing IFS

(Instruction Following Score)

The eval dataset in chat format

WRITER

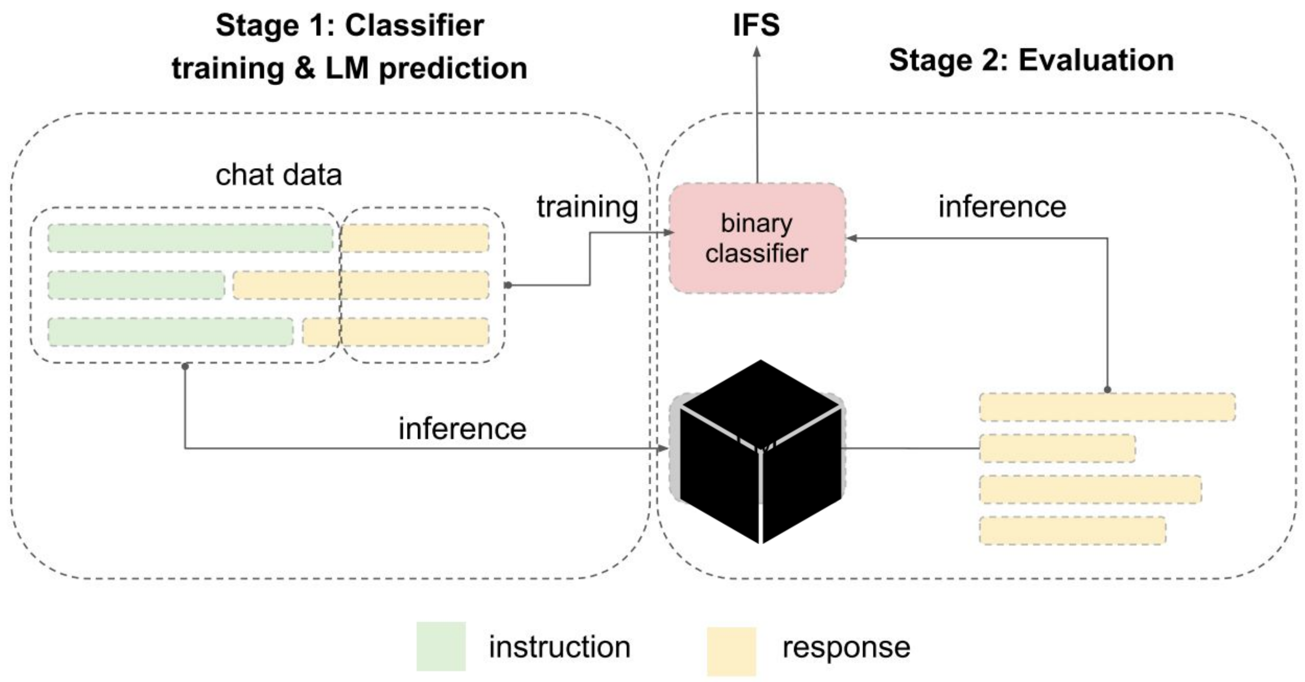


I → Instruction
R → Response
Ip → Instruction partial (fragmented instruction)
Ic → Continuation of instruction

Use a classifier to determine if the response is an instruct model or not

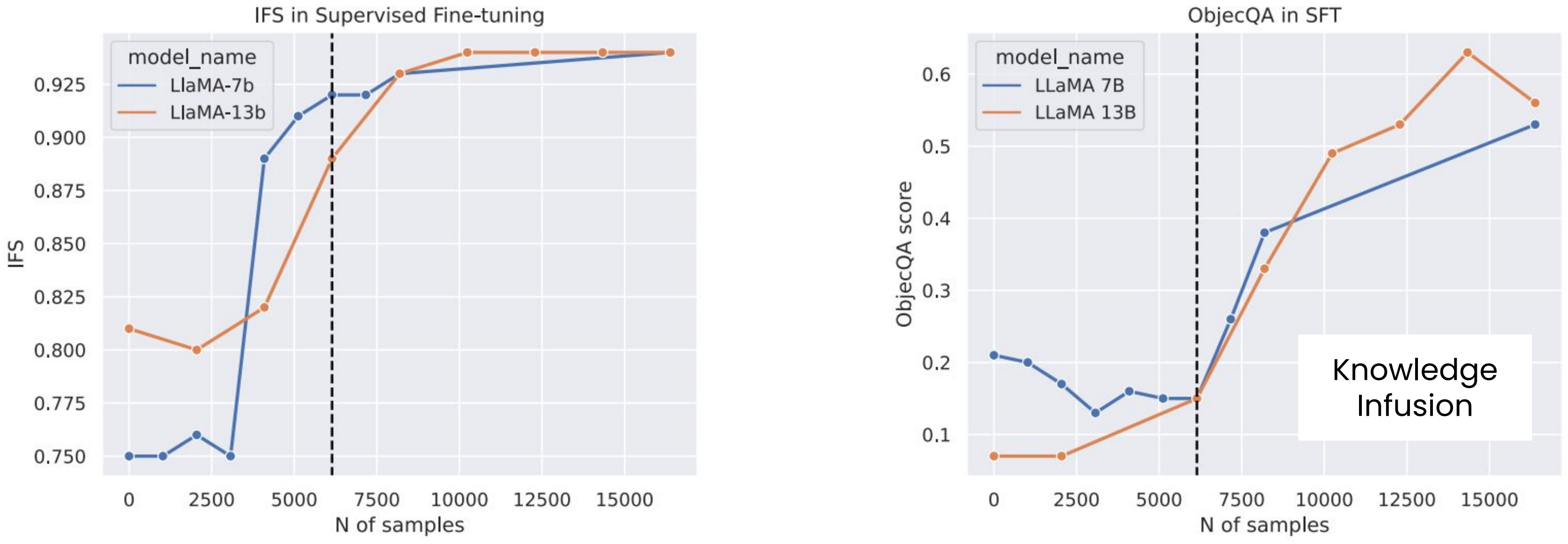
IFS experiment pipeline

WRITER



IFS as a stopping criterion in SFT

WRITER



(a) IFS (b) ObjecQA

Skills phase vs knowledge phase

WRITER

IFS can help you measuring specifics

Skills

- Text Generation
- Translation
- Question Answering
- Summarization
- Sentiment Analysis
- ...etc

Behaviors

- Bias
- Toxicity
- Creativity and Repetitiveness

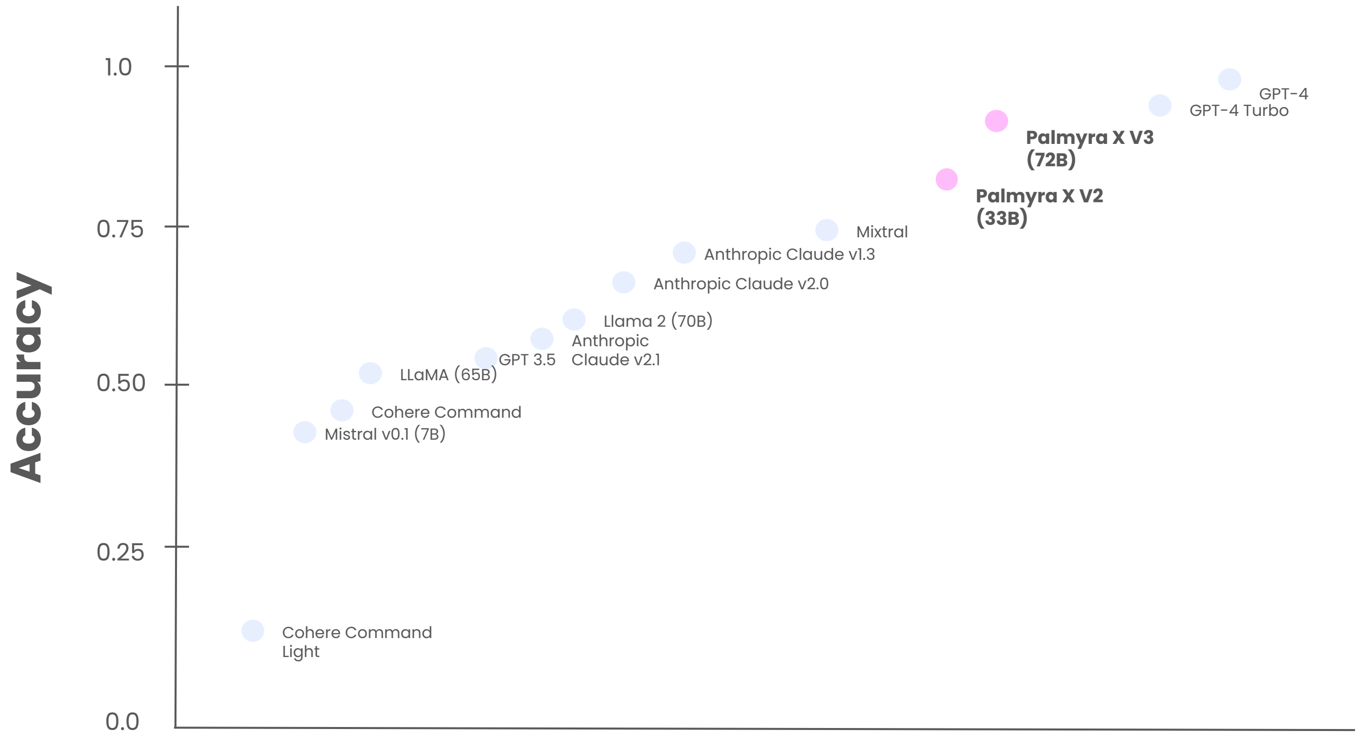
Knowledge

- Factual Knowledge
- Common Sense Knowledge
- Contextual Knowledge and Domain Specific Knowledge

So what?!

WRITER

Writer model accuracy vs other leading models



Release: v1.0.0 : 2023-12-19

GPT-4	0.962
GPT-4 Turbo	0.834
Palmyra X V3	**0.821**
Palmyra X V2	**0.783**
Mixtral	0.728
Anthropic Claude v1.3	0.724
Anthropic Claude 2.0	0.679
Llama 2 (70B)	0.659
GPT-3.5	0.621
Anthropic Claude 2.1	0.593
LLaMA (65B)	0.503
Cohere Command	0.462
Mistral v0.1 (7B)	0.438
Cohere Command Light	0.148

Source: https://crfm.stanford.edu/helm/lite/latest/#/leaderboard

< Papers | arxiv:2307.03692

Becoming self-instruct: introducing early stopping criteria for minimal instruct tuning

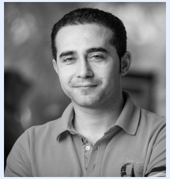
Published on Jul 5 · Featured in Daily Papers on Jul 10

Authors: Waseem AlShikh, Manhal Daaboul, Kirk Goddard, Brock Imel, Kiran Kamble, Parikshith Kulkarni, Melisa Russak

Abstract

In this paper, we introduce the Instruction Following Score (IFS), a metric that detects language models' ability to follow instructions. The metric has a dual purpose. First, IFS can be used to distinguish between base and instruct models. We benchmark publicly available base and instruct models, and show that the ratio of well formatted responses to partial and full sentences can be an effective measure between those two model classes. Secondly, the metric can be used as an early stopping criteria for instruct tuning. We compute IFS for Supervised Fine-Tuning (SFT) of 7B and 13B LLaMA models, showing that models learn to follow instructions relatively early in the training process, and the further finetuning can result in changes in the underlying base model semantics. As an example of semantics change we show the objectivity of model predictions, as defined by an auxiliary metric ObjecQA. We show that in this particular case, semantic changes are the steepest when the IFS tends to plateau. We hope that decomposing instruct tuning into IFS and semantic factors starts a new trend in better controllable instruct tuning and opens possibilities for designing minimal instruct interfaces querying foundation models.

The team:




Our paper

Q&A