DyKnow: Dynamically Verifying Time-Sensitive Factual Knowledge in LLMs

S. Mahed Mousavi, **Simone** Alghisi, Giuseppe Riccardi

Signals and Interactive Systems Lab, University of Trento, Italy

Introduction

LLMs are transforming to a ***common go-to*** for asking a vast variety of **factual questions**.

BUT! LLMs are trained on **static** data & evaluated via **static** benchmarks

Static benchmarks get **outdated** while factual knowledge is generally subject to **change**

DyKnow

We present **DyKnow** an evaluation framework to **dynamically** evaluate the knowledge in LLMs

For each fact, most **current info** are retrieved from the **Wikidata** at the time of evaluation

We evaluate the **24 LLMs** & **4 knowledge editing algorithms** regarding real-world facts.

Findings

LLMs **differ from traditional knowledge repositories**, making it important to investigate:

1. the **types of knowledge** these models can reliably manage
2. the **types of querying/alignment operations** they support

We **encourage** community **engagement** to **expand DyKnow into a current and active benchmark.**

RQ1. Accuracy & Consistency of SOTA LLM via DyKnow

Significant percentage of **outdated** or **irrelevant** responses in **all models**

High percentage of outdated responses **even in recent models**

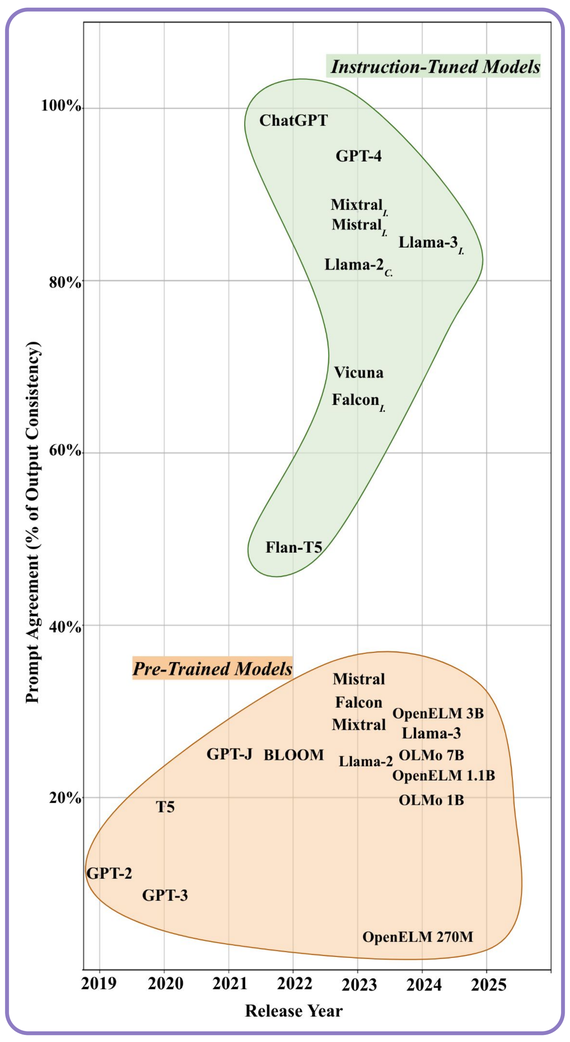
No model exceeds 80% correctness

Output **consistency**

1. **varies significantly** across different models
2. increases with **recent** models
3. is higher with Instruction-tuned models than pre-trained models

No model achieves 100% consistency

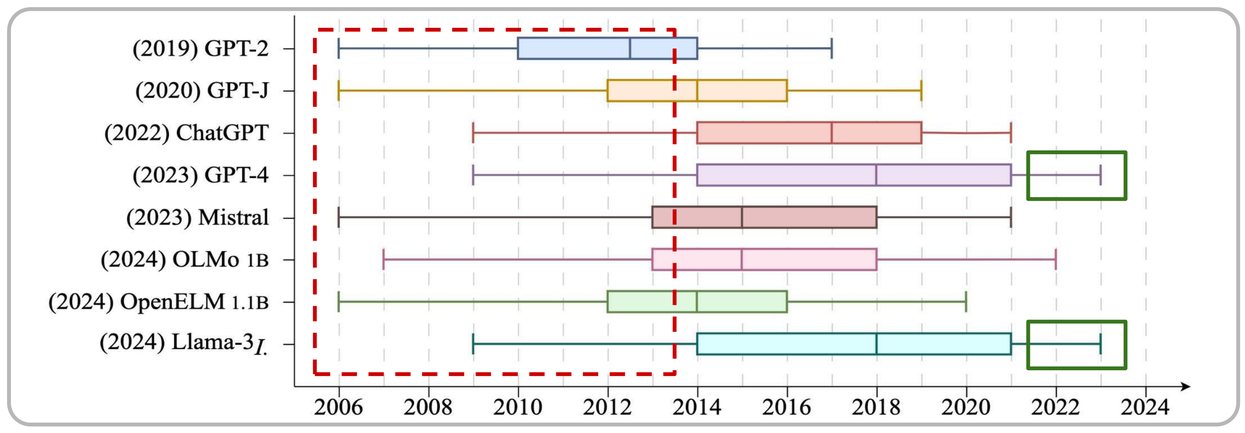
(Year) Model	Correct	Outdated	Irrelevant
(2019) GPT-2	26%	42%	32%
(2020) GPT-3	42%	47%	12%
(2020) T5	11%	21%	68%
(2021) GPT-J	41%	46%	13%
(2022) Bloom	35%	49%	16%
(2022) Flan-T5	18%	39%	43%
(2023) Llama-2	51%	42%	7%
(2023) Falcon	42%	47%	11%
(2023) Mistral	53%	39%	8%
(2023) Mixtral	48%	42%	10%
(2024) OLMo 1B	37%	40%	23%
(2024) OLMo 7B	35%	36%	29%
(2024) Llama-3	57%	36%	7%
(2024) OpenELM 270M	12%	28%	61%
(2024) OpenELM 1.1B	35%	47%	18%
(2024) OpenELM 3B	42%	42%	16%
(2022) ChatGPT	57%	35%	8%
(2023) GPT-4	80%	13%	7%
(2023) Llama-2$_{C.}$	51%	37%	12%
(2023) Falcon$_{I.}$	44%	41%	15%
(2023) Vicuna	52%	33%	15%
(2023) Mistral$_{I.}$	52%	32%	16%
(2023) Mixtral$_{I.}$	62%	29%	9%
(2024) Llama-3$_{I.}$	76%	14%	10%



RQ2: LLMs' Data Approximation

Approximating the LLMs' Pre-Training data shows:

1. **old information** across all models
2. **new information** in recent models
3. a trend of improvement in GPT model family



RQ3: Updating LLMs' Knowledge

SERAC fails to achieve high performance with updating at best less than 40%

ROME demonstrates an overall poor performance & is outperformed by **MEMIT**

MEMIT excels with GPT-J and Llama-2$_{C.}$ while **IKE** shows high success rates with *GPT-J* and Mistral$_{I.}$

Model	#Outdated Facts	Knowledge Editing			
		Modifying Parameters		Preserving Parameters	
		ROME	MEMIT	SERAC	IKE
(2019) GPT-2	54	17%	33%	4%	49%
(2021) GPT-J	60	11%	83%	0%	97%‡
(2023) Llama-2$_{C.}$	48	4%	77%	36%	18%
(2023) Mistral$_{I.}$	41	0%	0%	—	92%‡