universität wien

EMNLP 2024 Findings

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

To Know or Not To Know? Analyzing Self-Consistency of Large Language Models under Ambiguity

Anastasiia Sedova*, Robert Litschko*, Diego Frassinelli, Benjamin Roth, Barbara Plank

Poster: Nov 14 (Thursday) 10:30-12:00

Motivation

- Lack of self-consistency in LLMs ⇨ doubts about their trustworthiness and reliability
- …especially under ambiguity

- We conduct a behavioral study
- Desantangle *knowing* from *applying knowledge*…
- … and analyze the model behavior when faced with **entity ambiguity**

Behavioral Study - Ambiguous Entities

Fruits: Apple, Fig, Mango, Kiwi, Papaya, Orange

Animals: Jaguar, Puma, Fox, Lynx , Penguin, Greyhound, Dove

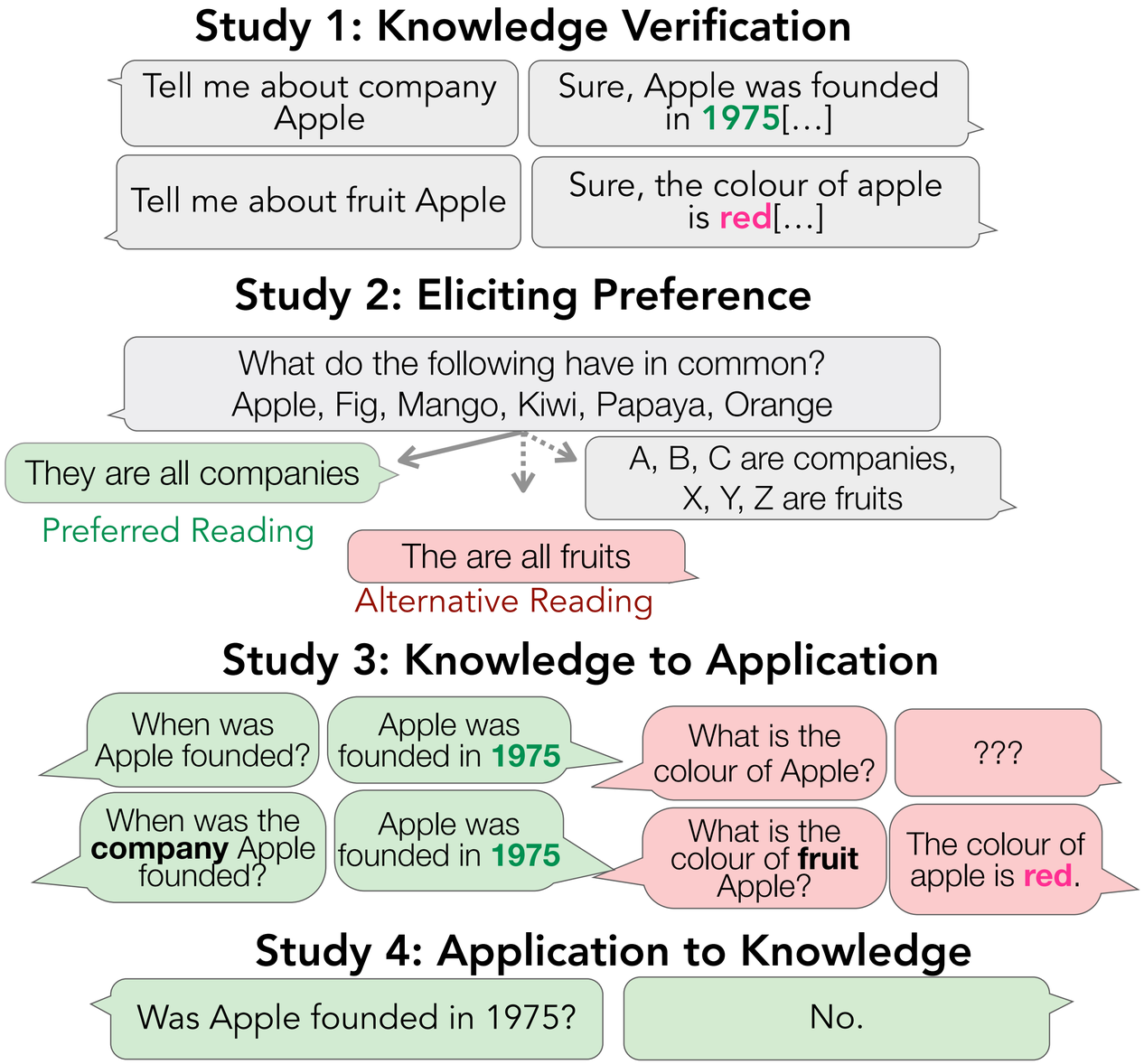
People: Ford, Disney, Tesla, Boeing, Dell, Ferrero, Benetton, Levi Strauss, Versace, Philips

Locations: Amazon, Cisco, Montblanc, Patagonia, Hershey, Nokia, Eagle Creek, Prosper

Myths: Amazon, Nike, Midas, Hyperion, Mars, Pegasus, Vulcan, Hermes

Abstract Concepts: Triumph, Harmony, Genesis, Vision, Pioneer, Vanguard, Zenith, Allure, Tempo, Fidelity

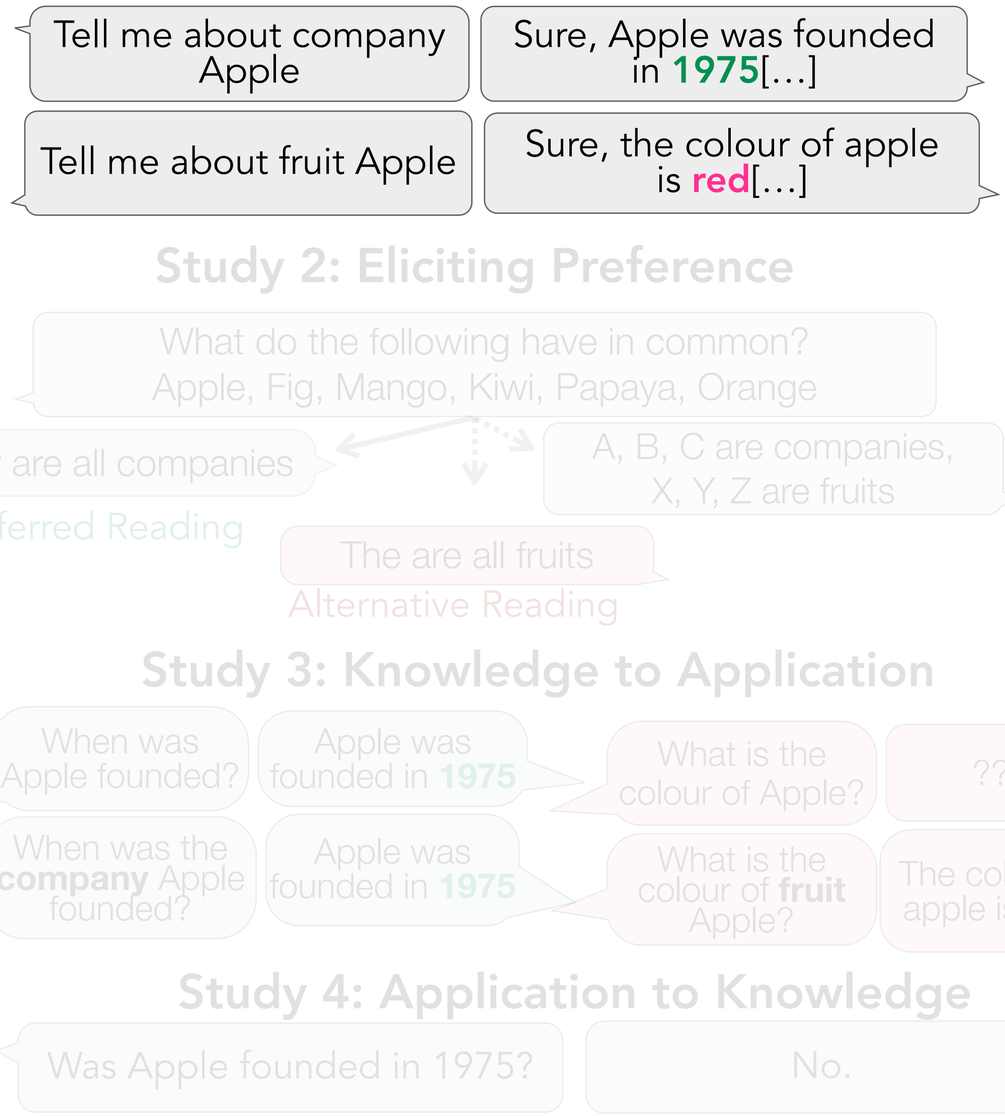
Behavioral Study - our 4 Studies



Behavioral Study - 1

- Sanity check
- All analyzed models are *aware* of both readings for all entities
- … but mostly failed to confirm the entity ambiguity:

Gemma	100.0	100.0	37.5	0.0	12.5	10.0
Mistral	100.0	83.8	75.0	10.0	75.0	90.0
Mixtral	71.4	50.0	0.0	0.0	30.0	50.0
GPT-3.5	57.1	100.0	0.0	10.0	12.5	10.0
GPT-4o	100.0	100.0	100.0	60.0	100.0	90.0
Llama-3	100.0	100.0	100.0	100.0	100.0	100.0



Behavioral Study - 2

- What is model's *preferred reading* of each entity type?
- More varied preferred readings for *Myths* and *Abstract* entities - possibly due to their higher ambiguity

Study 1: Knowledge Verification

Tell me about company Apple

Sure, Apple was founded in 1975[…]

Tell me about fruit Apple

Sure, the colour of apple is red[…]

Study 2: Eliciting Preference

What do the following have in common?
Apple, Fig, Mango, Kiwi, Papaya, Orange

They are all companies

Preferred Reading

A, B, C are companies, X, Y, Z are fruits

The are all fruits

Alternative Reading

Study 3: Knowledge to Application

When was Apple founded?

Apple was founded in 1975

What is the colour of Apple?

???

When was the **company** Apple founded?

Apple was founded in 1975

What is the colour of **fruit** Apple?

The colour of apple is red.

Study 4: Application to Knowledge

Was Apple founded in 1975?

No.

Behavioral Study - 3

How well can LLMs adopt the correct reading?

	Preferred Reading		Alternative Reading		Average		
	prop X	prop type X	prop X	prop type X	prop X	prop type X	Agg
Gemma	87.8	95.9	63.3	69.4	75.6	82.7	77.6
Mistral	77.6	100.0	63.3	87.8	70.5	93.9	82.2
Mixtral	77.6	100.0	75.5	85.7	76.6	92.9	84.8
GPT-3.5	87.8	100.0	75.5	77.6	81.7	88.8	85.3
GPT-4o	93.9	100.0	83.7	89.8	88.8	94.9	91.9
Llama-3	87.8	98.0	85.7	100.0	86.8	99.0	89.9
Average	85.4	99.0	74.5	85.1	80.0	90.5	85.3

Correlation with the entity popularity:

(Mixtral)
"Provide the gender for…

- "… Hermes'' -> "Hermes is a male deity in Greek mythology. […]"
- "…Amazon'' ->''Amazon.com, Inc. is a company, and as such, it does not have a gender. […]''

Study 1: Knowledge Verification

Tell me about company Apple

Sure, Apple was founded in 1975[…]

Tell me about fruit Apple

Sure, the colour of apple is red[…]

Study 2: Eliciting Preference

What do the following have in common? Apple, Fig, Mango, Kiwi, Papaya, Orange

They are all companies

A, B, C are companies, X, Y, Z are fruits

Preferred Reading

The are all fruits

Alternative Reading

Study 3: Knowledge to Application

When was Apple founded?

Apple was founded in **1975**

What is the colour of Apple?

???

When was the **company** Apple founded?

Apple was founded in **1975**

What is the colour of **fruit** Apple?

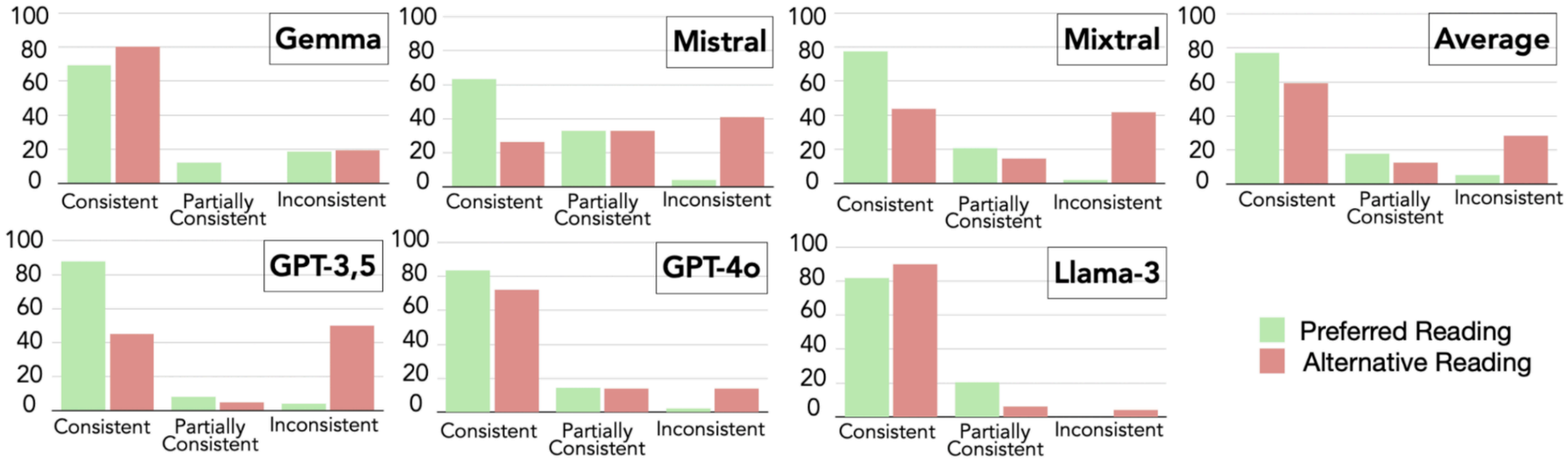
The colour of apple is **red**.

Study 4: Application to Knowledge

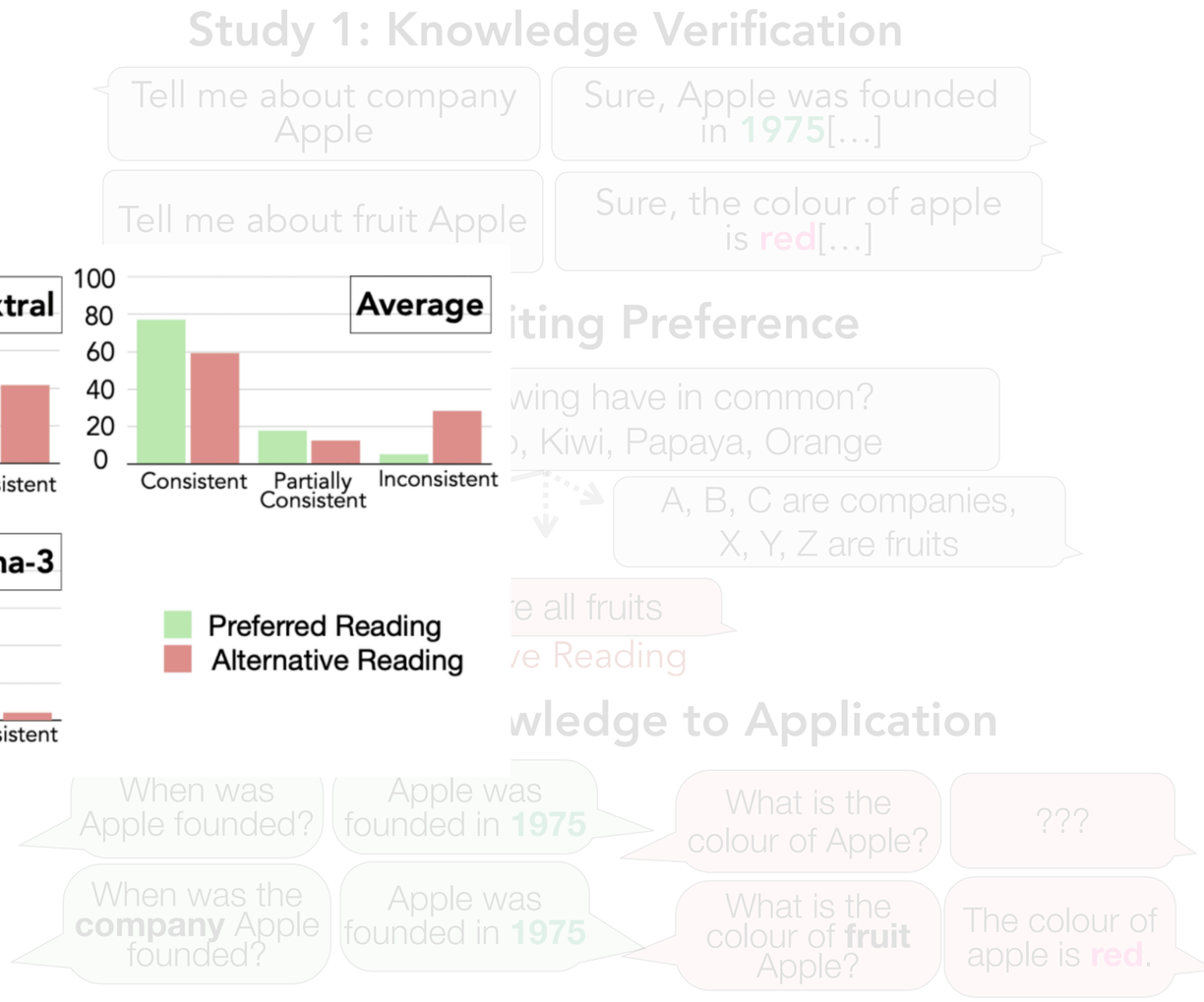
Was Apple founded in 1975?

No.

Behavioral Study - 4



(GPT-3.5) “No. December 5, 1901 is not the date of birth of Walt Disney. Walt Disney was actually born on December 5, 1901.”



Take Away

- LLMs often struggle to resolve entity ambiguity and correctly apply the knowledge they possess
- They exhibit biases toward preferred interpretations, influenced by the popularity of certain entities
- LLMs lack the ability to self-verify the accuracy of the knowledge they provide

See you at the poster! Nov 14 (Thursday) 10:30-12:00