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

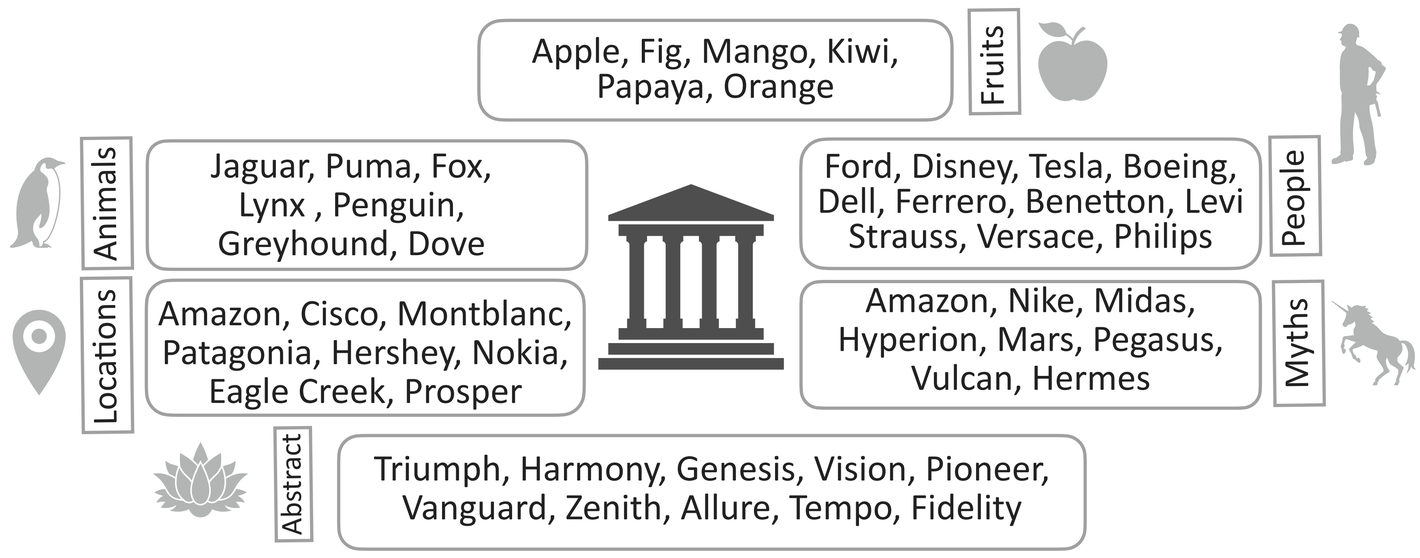
Anastasiia Sedova*[1], Robert Litschko*[2], Diego Frassinelli[2], Benjamin Roth[1], Barbara Plank[2]

*Equal Contribution [1] University of Vienna [2] LMU Munich

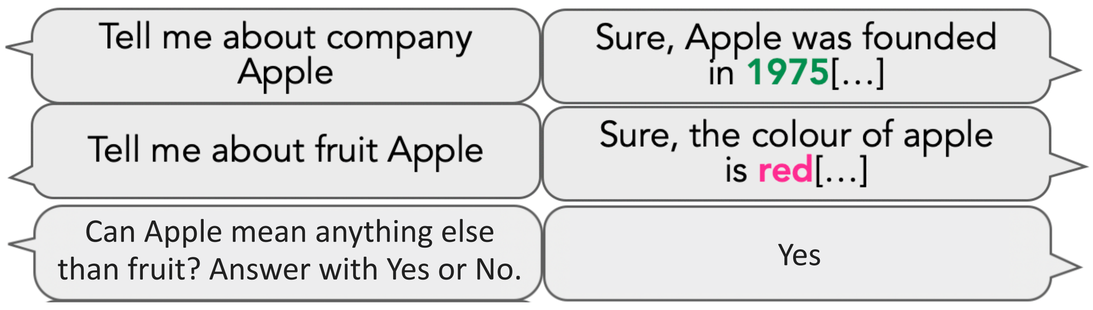
TLDR: SOTA LLMs fail to consistently apply factual knowledge under entity ambiguity

- Can LLMs implicitly resolve entity ambiguity?
- Are they capable of correctly applying their internal knowledge in ambiguous situations?
- How consistent are they in doing so? (⇨ **trustworthiness and reliability concerns**)

A behavioral test suite to analyze the LLMs behavior under entity ambiguity



Study 1: Knowledge Verification



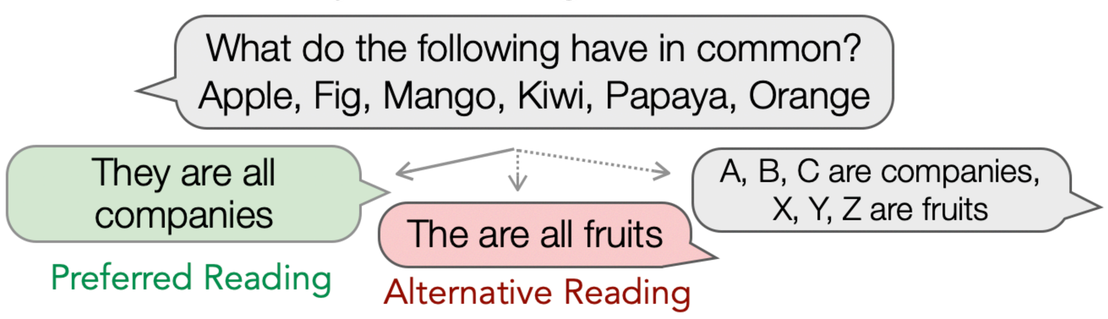
aka sanity check

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

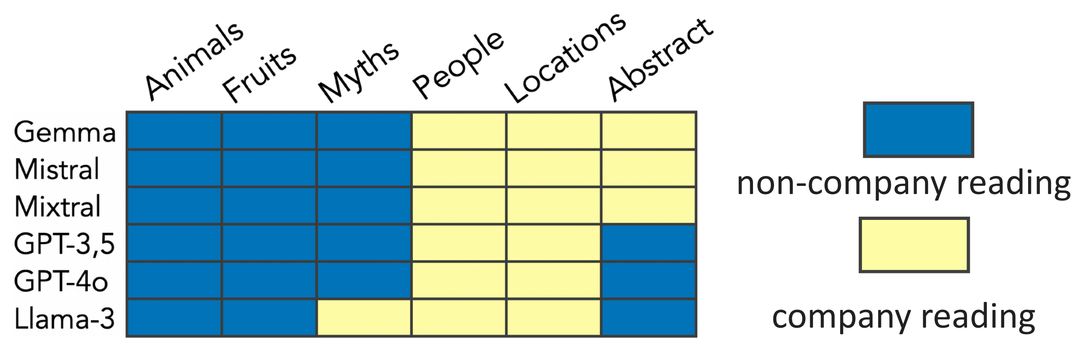
	Animals	Fruits	Myths	People	Locations	Abstract
Gemma	100.0	100.0	37.5	0.0	12.5	10.0
Mistral	100.0	83.8	75.0	10.0	75.0	90.0
Mixtral	71.4	50.0	0.0	0.0	30.0	50.0
GPT-3.5	57.1	100.0	0.0	10.0	12.5	10.0
GPT-4o	100.0	100.0	100.0	60.0	100.0	90.0
Llama-3	100.0	100.0	100.0	100.0	100.0	100.0

(The percentage of entities for which the models confirm ambiguity is reported.)

Study 2: Eliciting Preference

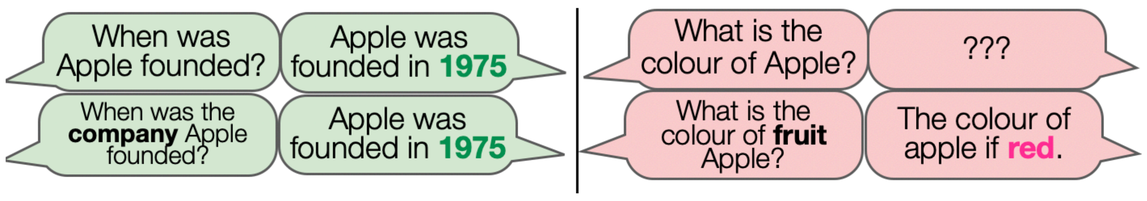


What is model's *preferred reading* of each entity type?



- More varied preferred readings for *Myths* and *Abstract* entities - possibly due to their higher ambiguity

Study 3: Knowledge to Application



How well can LLMs adopt the correct reading?

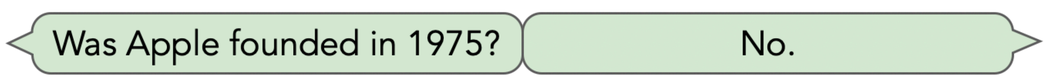
	Preferred Reading		Alternative Reading		Average		
	prop X	prop type X	prop X	prop type X	prop X	prop type X	Agg
Gemma	87.8	95.9	63.3	69.4	75.6	82.7	77.6
Mistral	77.6	100.0	63.3	87.8	70.5	93.9	82.2
Mixtral	77.6	100.0	75.5	85.7	76.6	92.9	84.8
GPT-3.5	87.8	100.0	75.5	77.6	81.7	88.8	85.3
GPT-4o	93.9	100.0	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

(The percentage of responses in which models adopted the correct interpretation is reported.)

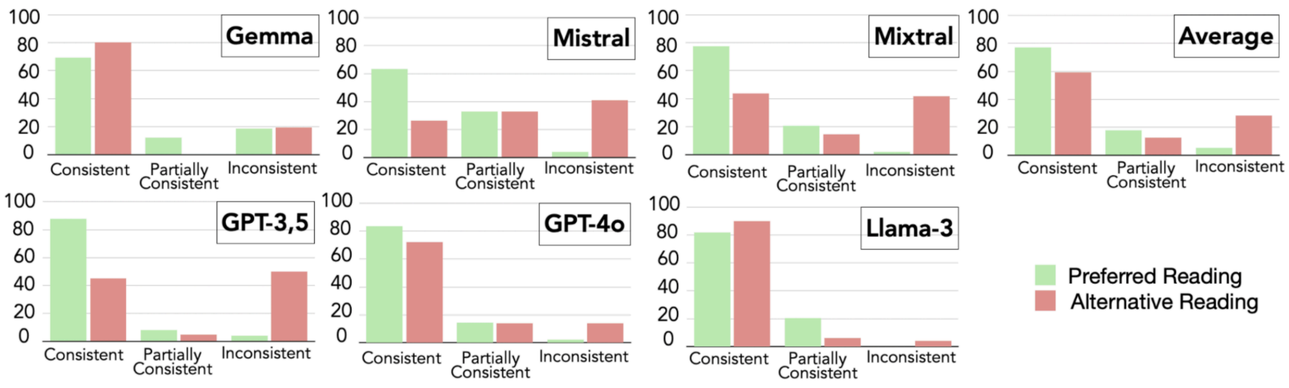
Correlation with the entity popularity:

(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 4: Application to Knowledge



Can LLMs reconfirm their knowledge?



(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."