

AnaloBench: Benchmarking the Identification of Abstract and Long-context Analogies

Xiao Ye, **Andrew Wang**,
Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala,
Nicholas Andrews, Daniel Khashabi



What makes two texts analogous?

The weather finally became pleasant following the stormy week.



As the flame extinguished, it left behind a thin wisp of smoke



What makes two texts analogous?

The weather finally became pleasant following the stormy week.

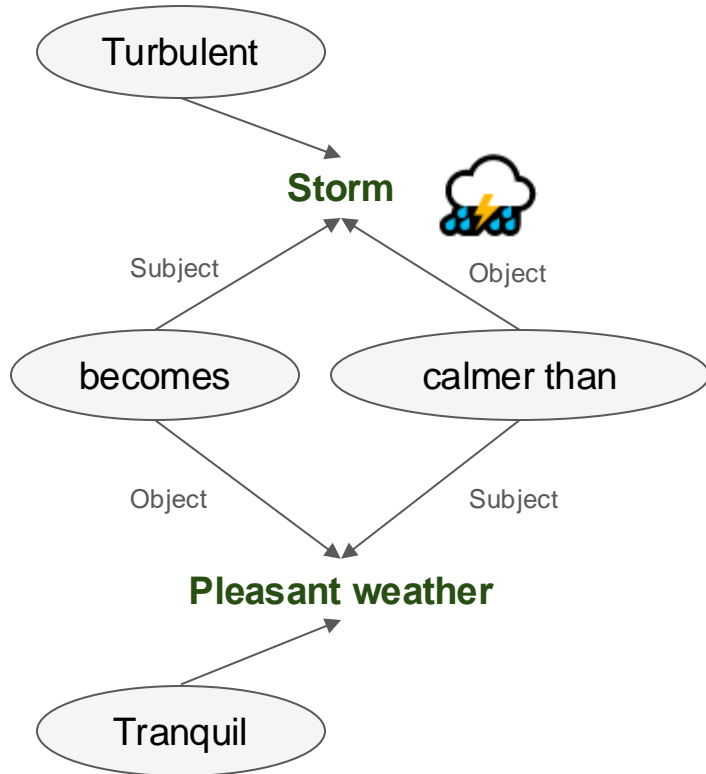


As the flame extinguished, it left behind a thin wisp of smoke

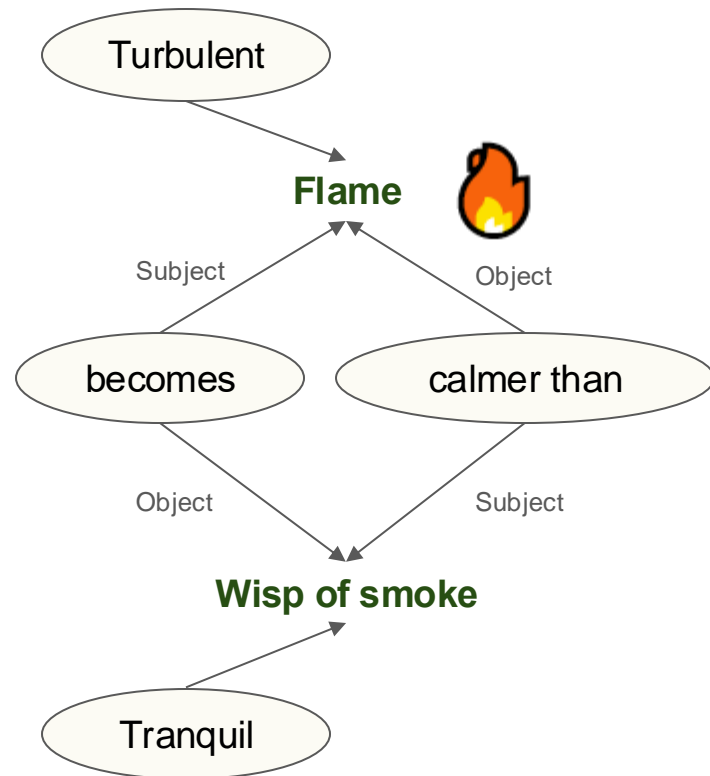
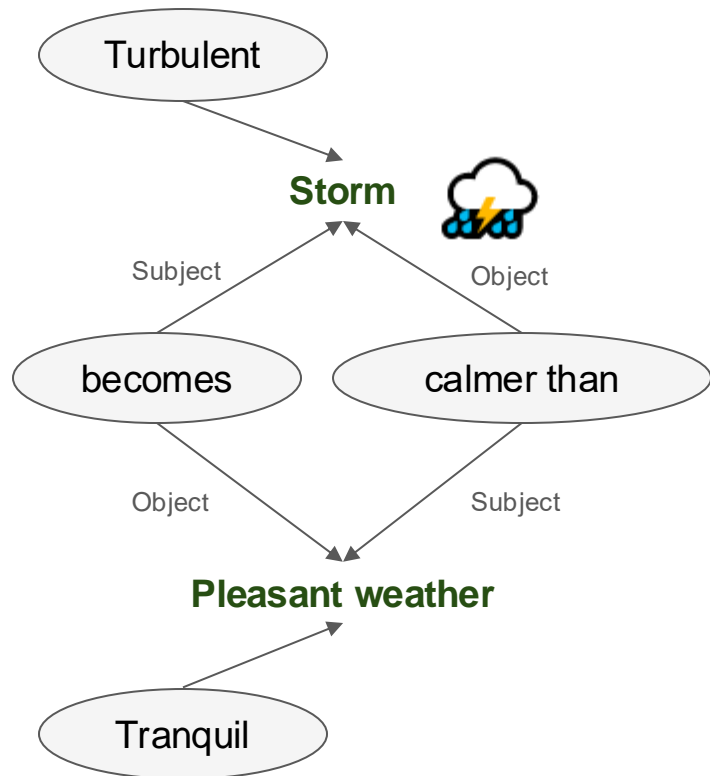


Analogous?

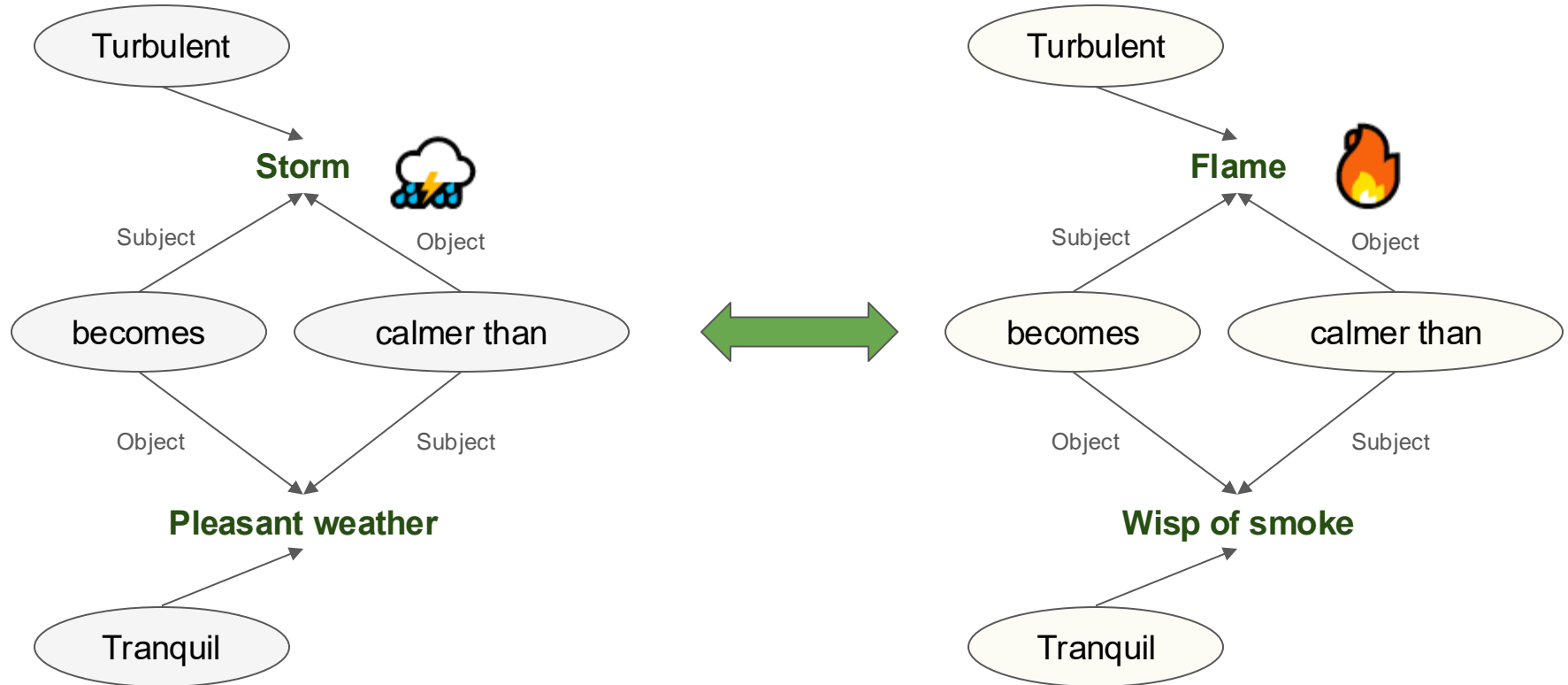
What makes two texts analogous?



What makes two texts analogous?



What makes two texts analogous?



What makes two texts analogous?

Turbulent

Turbulent

Structure Mapping Theory:

“An analogy is an assertion that a relational structure that normally applies in one domain can be applied in another domain”

Pleasant weather

Wisp of smoke

Tranquil

Structure-mapping: A theoretical framework for analogy
Dedre Gentner
Cognitive Science, 1983

Tranquil

Analogical reasoning in traditional AI systems

Challenge: how to automatically discover relational structures?

- 1980** — Patrick H Winston. **Learning and reasoning by analogy.** Communications of the ACM.
- 1983** — Jaime G Carbonell. **Learning by analogy: Formulating and generalizing plans from past experience.** In Machine learning.
- 1984** — Douglas R Hofstadter. **The copycat project: An experiment in nondeterminism and creative analogies.**
- 1999** — Roger C Schank. **Dynamic memory revisited.**
- 2013** — Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. **Linguistic regularities in continuous space word representations.** NAACL

Analogical reasoning in traditional AI systems

Challenge: how to automatically discover relational structures?

- 1980** — Patrick H Winston. **Learning and reasoning by analogy.** Communications of the ACM.
- 1983** — Jaime G Carbonell. **Learning by analogy: Formulating and generalizing plans from past experience.** In Machine learning.
- 1984** — Douglas R Hofstadter. **The copycat project: An experiment in nondeterminism and creative analogies.**
- 1999** — Roger C Schank. **Dynamic memory revisited.**
- 2013** — Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. **Linguistic regularities in continuous space word representations.** NAACL

LLMs and analogical reasoning

2023

Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

2024

Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors

Nicholas Ichien^{a,1}

Dušan Stamenković^b

Keith J. Holyoak^c

LLMs and analogical reasoning

Emergent Analogical Reasoning in Large Language Models

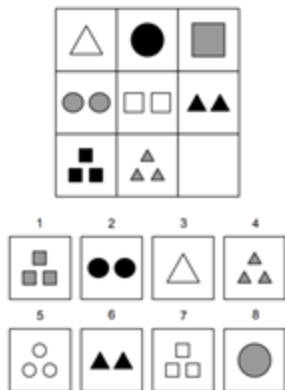
Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

“large language models...have acquired an emergent ability to find zero-shot solutions to a broad range of analogy problems”

LLMs and analogical reasoning

Emergent Analogical Reasoning in Large Language Models

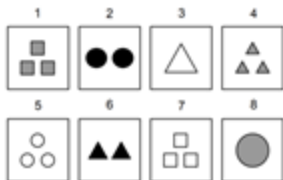
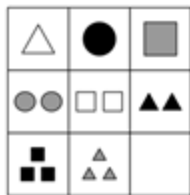
Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}



LLMs and analogical reasoning

Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}



Categorical

vegetable : cabbage :: insect : ?

1. beetle
2. frog

Function

drive : car :: burn : ?

1. wood
2. fire

Antonym

love : hate :: rich : ?

1. poor
2. wealthy

Synonym

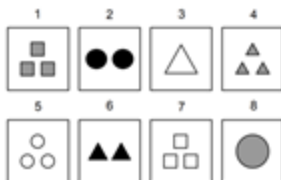
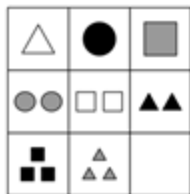
rob : steal :: cry : ?

1. weep
2. laugh

LLMs and analogical reasoning

Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}



Categorical
vegetable : cabbage :: insect : ?

1. beetle
2. frog

Function
drive : car :: burn : ?

1. wood
2. fire

Antonym
love : hate :: rich : ?

1. poor
2. wealthy

Synonym
rob : steal :: cry : ?

1. weep
2. laugh

Source story: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Far analogy – correct target story: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.

Towards more challenging evaluations

Real world analogy: the **solar system** is like the **atom**

☰ Solar System

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

For other uses, see [Solar System \(disambiguation\)](#).

The **Solar System**^[d] is the [gravitationally bound](#) system of the [Sun](#) and the objects that orbit it.^[11] It formed about 4.6 billion years ago when a dense region of a [molecular cloud](#) collapsed, forming the Sun and a [protoplanetary disc](#). The Sun is a typical star that maintains a [balanced equilibrium](#) by the [fusion](#) of hydrogen into helium at its [core](#), releasing this energy from its outer [photosphere](#). Astronomers [classify](#) it as a [G-type main-sequence star](#).

The largest objects that orbit the Sun are the eight [planets](#). In order from the Sun, they are four [terrestrial planets](#) ([Mercury](#), [Venus](#), [Earth](#) and [Mars](#)); two [gas giants](#) ([Jupiter](#) and [Saturn](#)); and two [ice giants](#) ([Uranus](#) and [Neptune](#)). All terrestrial planets have solid surfaces. Inversely, all [giant planets](#) do not have a definite surface, as they are mainly composed of gases and liquids. Over 99.86% of the Solar System's mass is in the Sun and nearly 90% of the remaining mass is in Jupiter and Saturn.

There is a strong consensus among astronomers^[e] that the Solar System has at least nine [dwarf planets](#): [Ceres](#), [Orcus](#), [Pluto](#), [Haumea](#), [Quaoar](#), [Makemake](#), [Gonggong](#), [Eris](#),

☰ Atom

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

For other uses, see [Atom \(disambiguation\)](#).

Atoms are the basic particles of the [chemical elements](#). An atom consists of a [nucleus](#) of [protons](#) and generally [neutrons](#), surrounded by an electromagnetically bound swarm of [electrons](#). The chemical elements are distinguished from each other by the number of protons that are in their atoms. For example, any atom that contains 11 protons is [sodium](#), and any atom that contains 29 protons is [copper](#). Atoms with the same number of protons but a different number of neutrons are called [isotopes](#) of the same element.

Atoms are extremely small, typically around 100 [picometers](#) across. A human hair is about a million carbon atoms wide. Atoms are smaller than the shortest wavelength of visible light, which means humans cannot see atoms with conventional microscopes. They are so small that accurately predicting their behavior using [classical physics](#) is not possible due to [quantum effects](#).

More than 99.94% of an atom's [mass](#) is in the nucleus. Protons have a positive [electric charge](#) and neutrons have no charge, so the nucleus is positively charged. The electrons are negatively charged, and this opposing charge is what binds them to the nucleus. If the numbers of [protons](#) and electrons are equal, as they normally are, then

Towards more challenging evaluations

Real world analogy: the solar system is like the atom

☰ Solar System

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

☰ Atom

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

Are LLMs performant on more challenging analogical reasoning tasks?

The largest objects that orbit the Sun are the eight [planets](#). In order from the Sun, they are four [terrestrial planets](#) ([Mercury](#), [Venus](#), [Earth](#) and [Mars](#)); two [gas giants](#) ([Jupiter](#) and [Saturn](#)); and two [ice giants](#) ([Uranus](#) and [Neptune](#)). All terrestrial planets have solid surfaces. Inversely, all [giant planets](#) do not have a definite surface, as they are mainly composed of gases and liquids. Over 99.86% of the Solar System's mass is in the Sun and nearly 90% of the remaining mass is in Jupiter and Saturn.

There is a strong consensus among astronomers^[e] that the Solar System has at least nine [dwarf planets](#): [Ceres](#), [Orcus](#), [Pluto](#), [Haumea](#), [Quaoar](#), [Makemake](#), [Gonggong](#), [Eris](#),

Atoms are extremely small, typically around 100 [picometers](#) across. A human hair is about a million carbon atoms wide. Atoms are smaller than the shortest wavelength of visible light, which means humans cannot see atoms with conventional microscopes. They are so small that accurately predicting their behavior using [classical physics](#) is not possible due to [quantum effects](#).

More than 99.94% of an atom's [mass](#) is in the nucleus. Protons have a positive [electric charge](#) and neutrons have no charge, so the nucleus is positively charged. The electrons are negatively charged, and this opposing charge is what binds them to the nucleus. If the numbers of [protons](#) and electrons are equal, as they normally are, then

Tasks

Task 1:
Multiple choice

Task 2:
Analogical retrieval

Dataset creation

340 seed analogies from human annotators

① Analogical Pairs

Sentence A	Sentence B
<i>You can't pour from an empty cup.</i>	<i>A fallen tree cannot provide shade.</i>
<i>You can't pour from an empty cup.</i>	<i>It's hard to love with a broken heart.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>After letting off his rage he sat down like a lamb.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>The weather finally became pleasant following the stormy week.</i>
⋮	⋮

Dataset creation

340 seed analogies from human annotators

① Analogical Pairs

Sentence A	Sentence B
<i>You can't pour from an empty cup.</i>	<i>A fallen tree cannot provide shade.</i>
<i>You can't pour from an empty cup.</i>	<i>It's hard to love with a broken heart.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>After letting off his rage he sat down like a lamb.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>The weather finally became pleasant following the stormy week.</i>
⋮	⋮

Each analogy is a pair of analogous sentences

Dataset creation

340 seed analogies from human annotators

① Analogical Pairs

Sentence A	Sentence B
<i>You can't pour from an empty cup.</i>	<i>A fallen tree cannot provide shade.</i>
<i>You can't pour from an empty cup.</i>	<i>It's hard to love with a broken heart.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>After letting off his rage, he sat down like a lamb.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>The weather finally became pleasant following the stormy week.</i>
⋮	⋮

annotated to be analogous

pre-existing sentences

Dataset creation

① Analogical Pairs

Sentence A	Sentence B
<i>You can't pour from an empty cup.</i>	<i>A fallen tree cannot provide shade.</i>
<i>You can't pour from an empty cup.</i>	<i>It's hard to love with a broken heart.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>After letting off his rage he sat down like a lamb.</i>
<i>He danced off his sugar high then promptly fell asleep.</i>	<i>The weather finally became pleasant following the stormy week.</i>
⋮	⋮

Analogous texts are grouped into clusters

② Analogical Clusters

- *After letting off his rage he sat down like a lamb.*
- *He danced off his sugar high then promptly fell asleep.*
- *The weather finally became pleasant following the stormy week.*

- *You can't pour from an empty cup.*
- *A fallen tree cannot provide shade.*
- *It's hard to love with a broken heart.*
- ...

⋮

Tasks

**Task 1:
Multiple choice**

**Task 2:
Analogical retrieval**

Task 1: Multiple choice

Description: Given a narrative, can LLMs identify the most analogous narrative from a set of 4 choices?

Q: Which of the following is the most analogous sentence to the target sentence?

Target: The weather finally became pleasant following the stormy week.

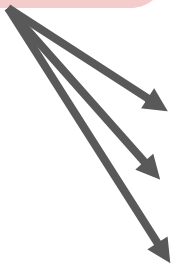
- ✓ **As the flame extinguished, it left behind a...**
- ✗ All that glitters is not gold.
- ✗ A fallen tree cannot provide shade.
- ✗ A succession of waves battered the rock.

Task 1: Multiple choice

Description: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?

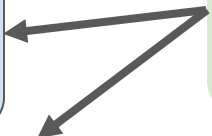
Q: Which of the following is the most analogous sentence to the target sentence?

Sampled from
outside of
cluster



Target: The weather finally became pleasant following the stormy week.

Sampled from
same cluster



- ✓ **As the flame extinguished, it left behind a...**
- ✗ All that glitters is not gold.
- ✗ A fallen tree cannot provide shade.
- ✗ A succession of waves battered the rock.

Task 1: Multiple choice

Intuition: analogies between longer narratives are more difficult

Q: Which of the following is the most analogous sentence to the target sentence?

Target: The weather finally became pleasant following the stormy week.

- ✓ **As the flame extinguished, it left behind a...**
- ✗ All that glitters is not gold.
- ✗ A fallen tree cannot provide shade.
- ✗ A succession of waves battered the rock.

Task 1: Multiple choice

Intuition: analogies between longer narratives are more difficult

Q: Which of the following is the most analogous story to the target story?

Target: For an entire week, the town had been battered by incessant, violent storms. The sky was a relentless shade of charcoal, thunder...

- ✓ **Once upon a time, in a quiet town nestled amidst snow-capped...**
- ✗ In the small town of Baker's Crest, people lived simple lives, mostly...
- ✗ Once upon a time, a man named Henry lived in a tiny, worn-out house. He...
- ✗ A stormy morning on the coast revealed an awe-inspiring sight as a ...

Lengthening seed analogies

GPT-4 prompting: Increase length of narratives in analogies to 10 sentences and 30 sentences

Prompt:

Given the following sentence, expand it into a 30 sentence story:

Sentence: **The weather finally became pleasant following the stormy week.**

For an entire week, the town had been battered by incessant, violent storms. The sky was a relentless shade of charcoal, thunder...

Quality of these extended stories is attested in Appendix C



GPT-4

Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?

Model ↓ - Story length →	1-sent	10-sent	30-sent	
Random	25	25	25	

Open-source	Zephyr (7B)	55.1	27.1	20.3
	UnifiedQA (11B)	68.1	27.3	17.8
	WizardLM (13B)	41.1	29.1	25.7
	LLaMA2-chat (70B)	55.6	39.2	29.5
	XwinLM (70B)	66.3	35.7	26.8
	Tulu2 (70B)	71.8	51.2	31.5

Private	Claude	68.2	30.2	25.9
	GPT3.5	65.3	46.4	30.8
	GPT4	89.1	66.5	60.7

Human	96.0	72.5	73.3	

Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?

Model ↓ - Story length →		1-sent	10-sent	30-sent
Random		25	25	25
Open-source	Zephyr (7B)	55.1	27.1	20.3
	UnifiedQA (11B)	68.1	27.3	17.8
	WizardLM (13B)	41.1	29.1	25.7
	LLaMA2-chat (70B)	55.6	39.2	29.5
	XwinLM (70B)	66.3	35.7	26.8
	Tulu2 (70B)	71.8	51.2	31.5
Private	Claude	68.2	30.2	25.9
	GPT3.5	65.3	46.4	30.8
	GPT4	89.1	66.5	60.7
Human		96.0	72.5	73.3

Human performance is high but decreases with story length

Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?

Model ↓ - Story length →	1-sent	10-sent	30-sent
Random	25	25	25

Open-source			
Zephyr (7B)	55.1	27.1	20.3
UnifiedQA (11B)	68.1	27.3	17.8
WizardLM (13B)	41.1	29.1	25.7
LLaMA2-chat (70B)	55.6	29.2	29.5
XwinLM (70B)	66.3	35.7	26.8
Tulu2 (70B)	71.8	51.2	31.5

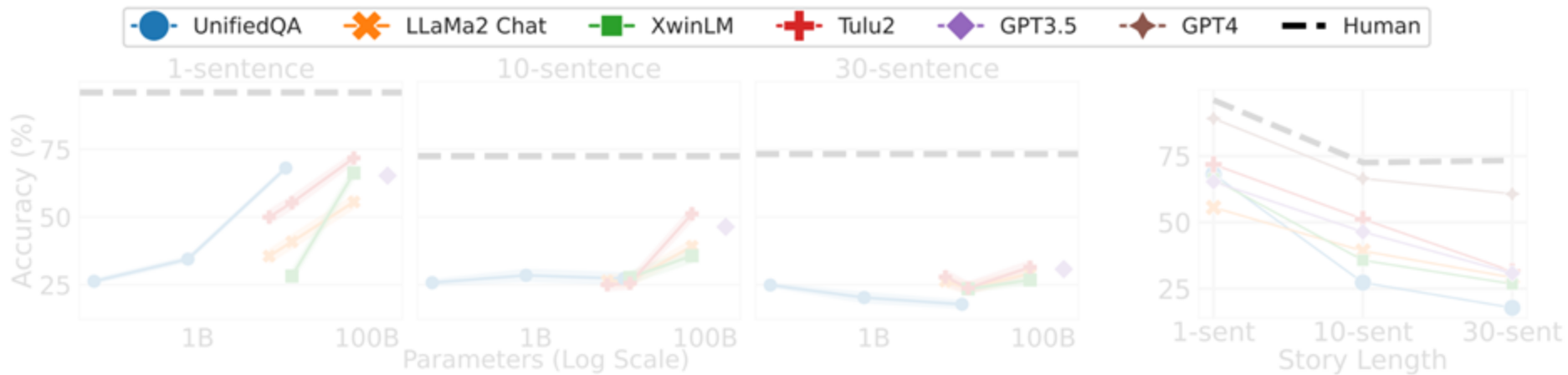
Private			
Claude	68.2	30.2	25.9
GPT3.5	65.3	46.4	30.8
GPT4	89.1	66.5	60.7

Human	96.0	72.5	73.3

LLM performance decrease
is much larger than
human performance decrease

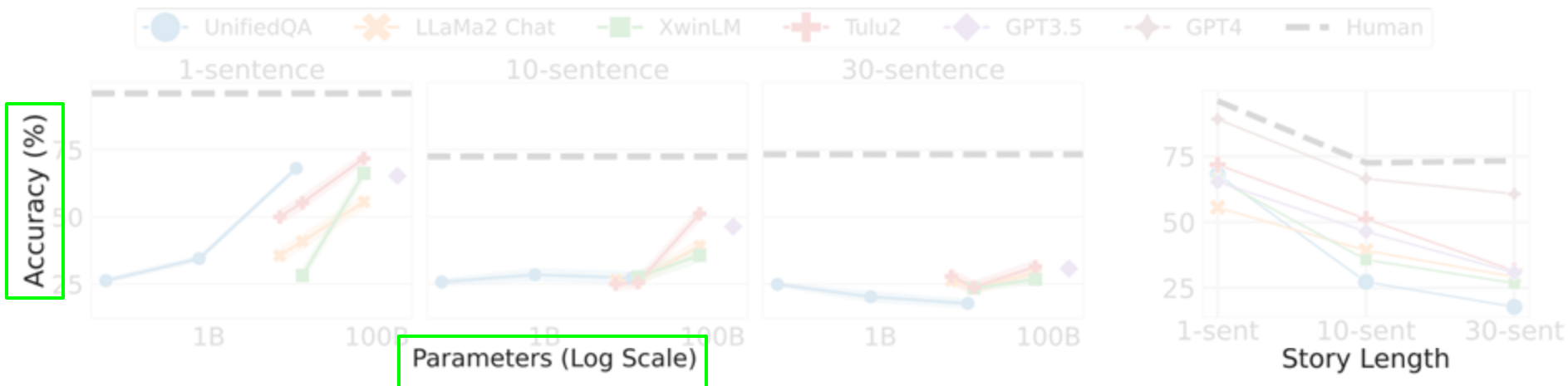
Further analysis - model size

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?



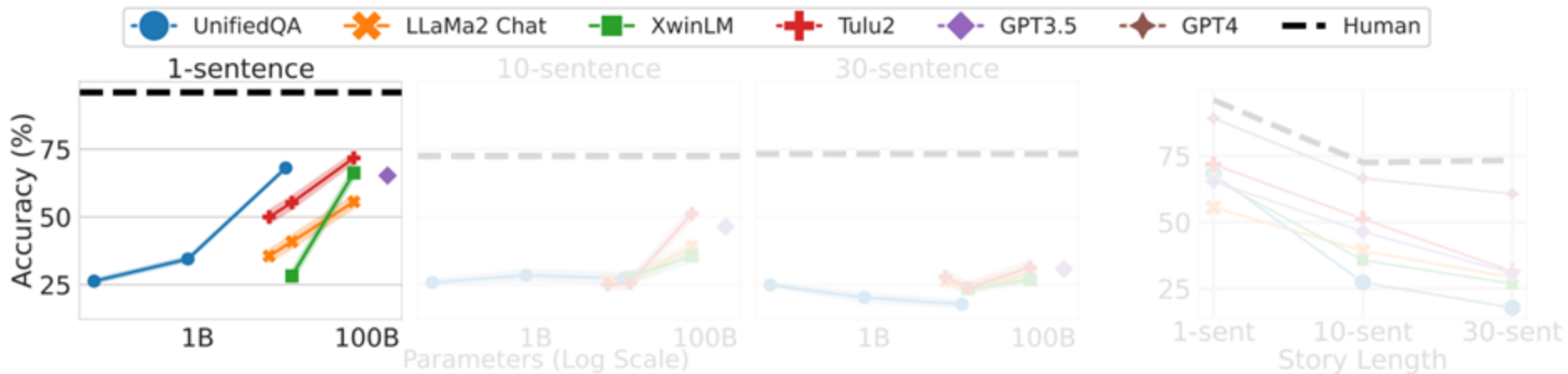
Further analysis - model size

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of 4 choices?



Further analysis - model size

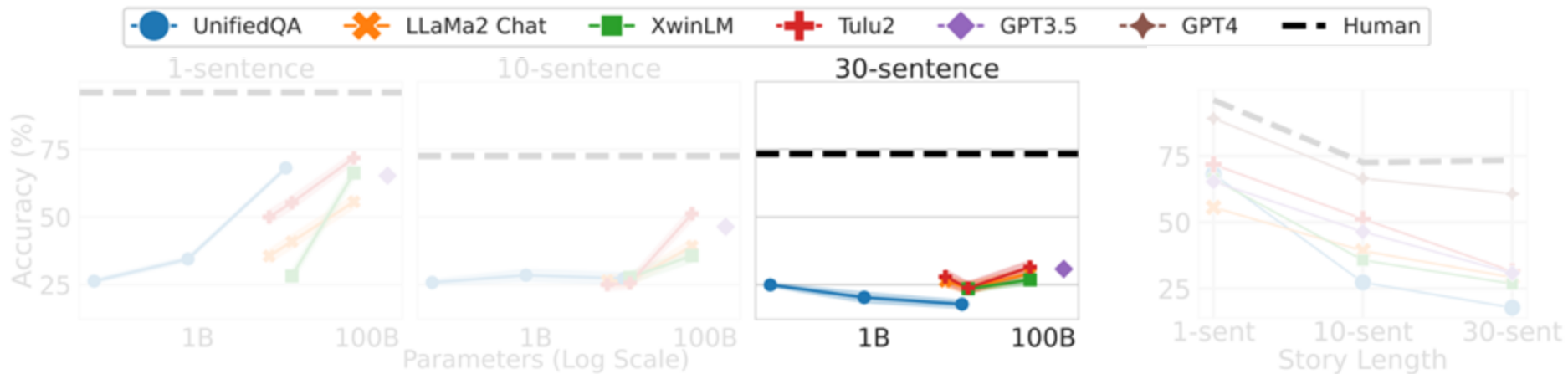
Task Overview: Effect of model size on performance



Performance scales with model size on short narratives

Further analysis - model size

Task Overview: Effect of model size on performance



Performance **DOES NOT** scale with model size on long narratives

Tasks

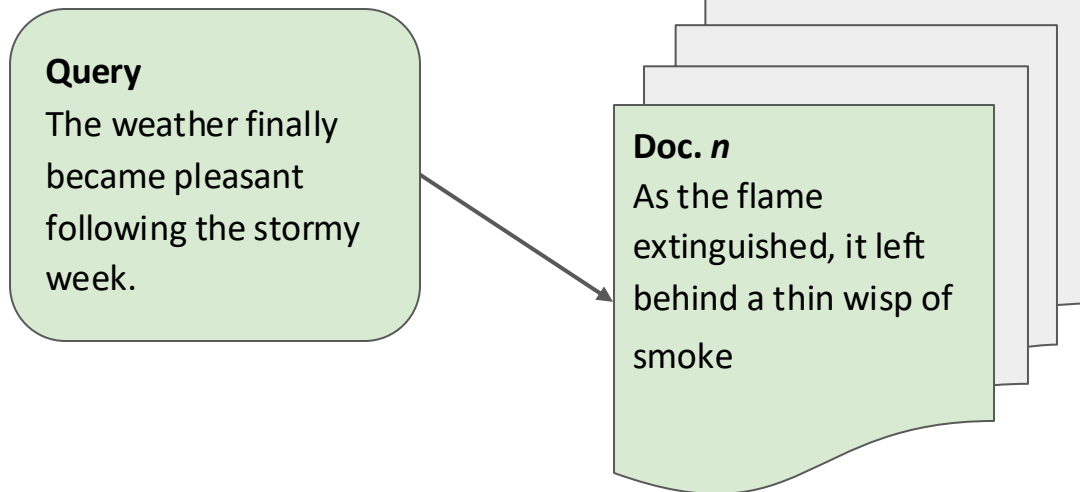
Task 1:
Challenging analogies

Task 2:
Analogical retrieval

Task 2 - Analogical retrieval

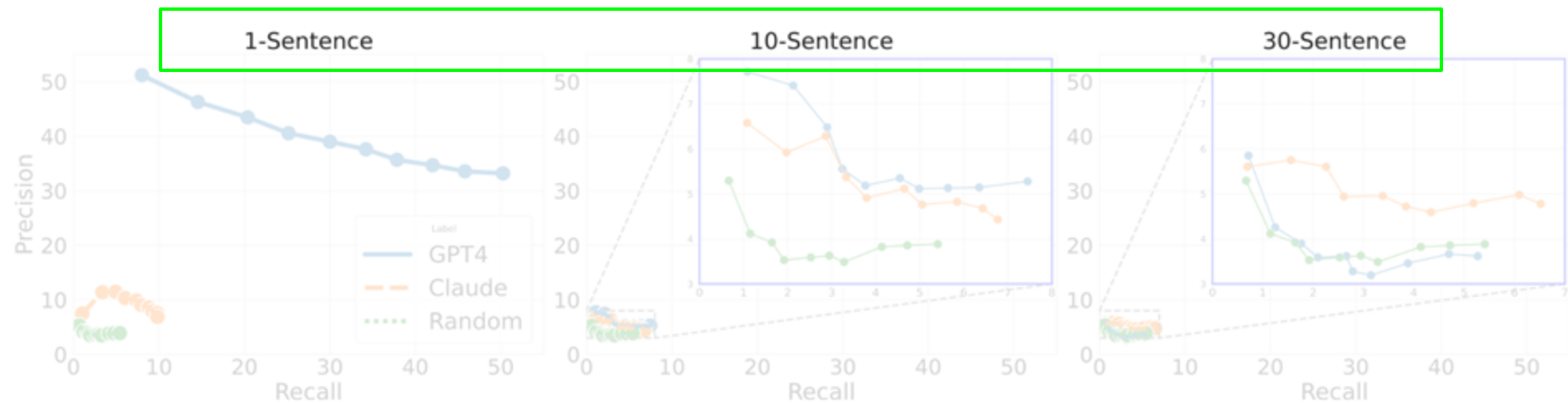
Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?

Q: Retrieve the top 10 analogous stories from the sentence bank...



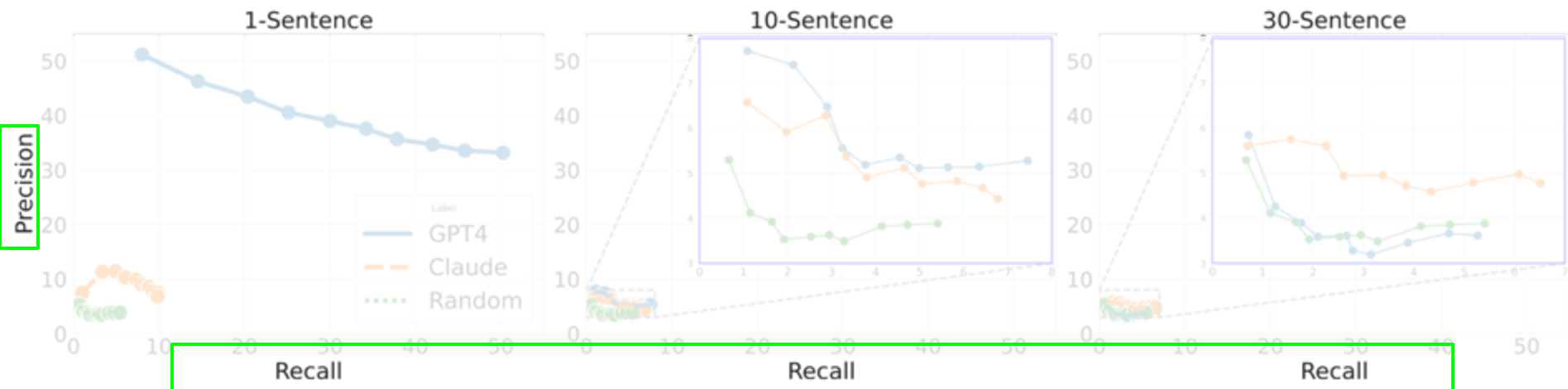
Task 2 - Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?



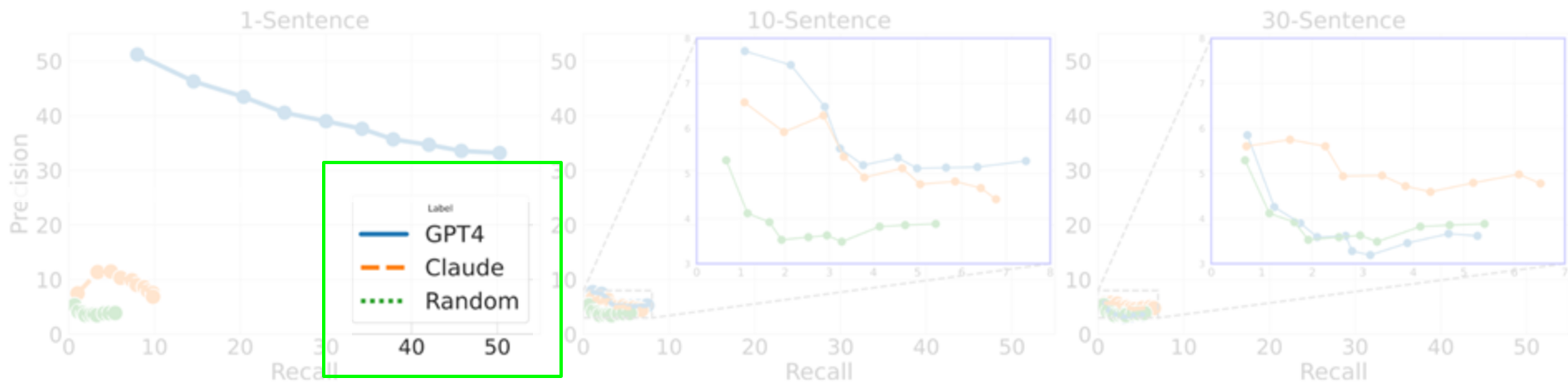
Task 2 - Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?



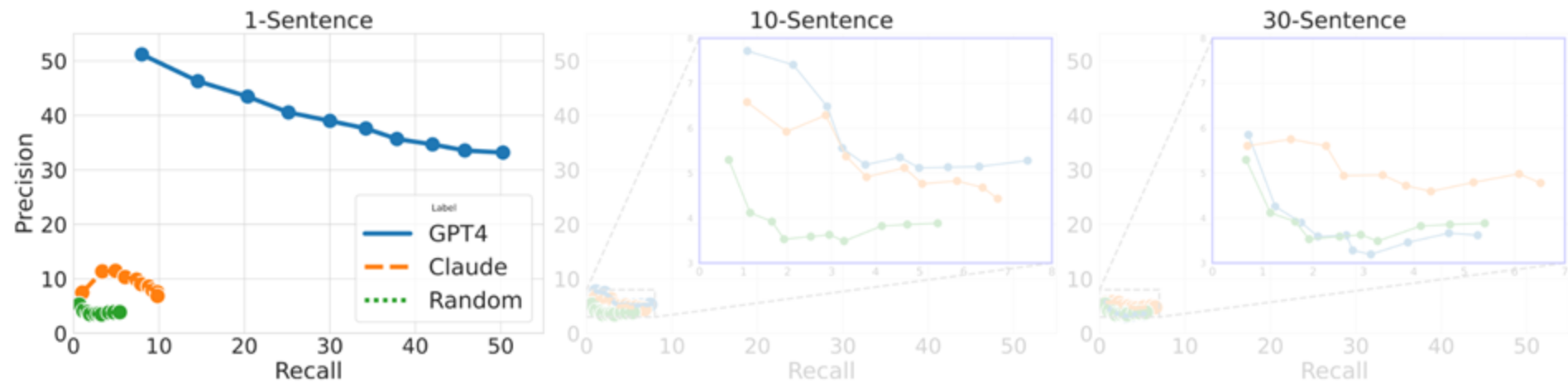
Task 2 - Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?



Task 2 - Results

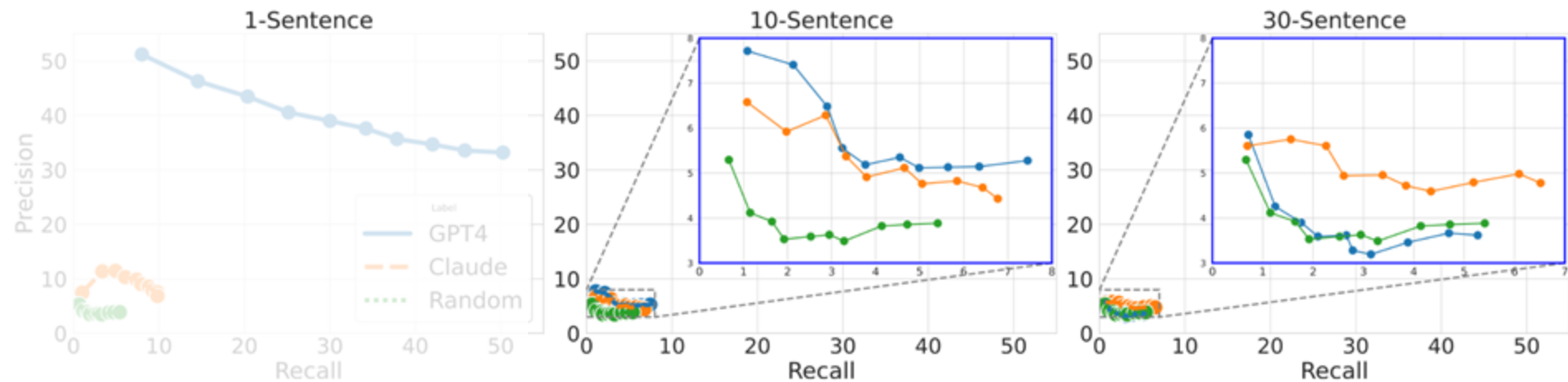
Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?



Some LLMs do well on short narratives

Task 2 - Results

Task Overview: Given a narrative, can LLMs identify the most analogous sentence from a set of **200** choices?



All LLMs perform trivially on long narratives

Conclusion

Are LLMs performant on more challenging analogical reasoning tasks?

Conclusion

Are LLMs performant on more challenging analogical reasoning tasks?

Task 1:
**Challenging analogies are
difficult for LLMs**

Human-AI ability gap
increases on longer narratives

Conclusion

Are LLMs performant on more challenging analogical reasoning tasks?

Task 1:
**Challenging analogies are
difficult for LLMs**

Model size does not help
when narratives are long

Conclusion

Are LLMs performant on more challenging analogical reasoning tasks?

Task 1:
**Challenging analogies are
difficult for LLMs**

Task 2:
**LLMs perform poorly on
analogical retrieval**

Performance on longer
narratives is trivial

Conclusion

Are LLMs performant on more challenging analogical reasoning tasks?

Task 1:
**Challenging analogies are
difficult for LLMs**

Task 2:
**LLMs perform poorly on
analogical retrieval**

AnaloBench is challenging for LLMs