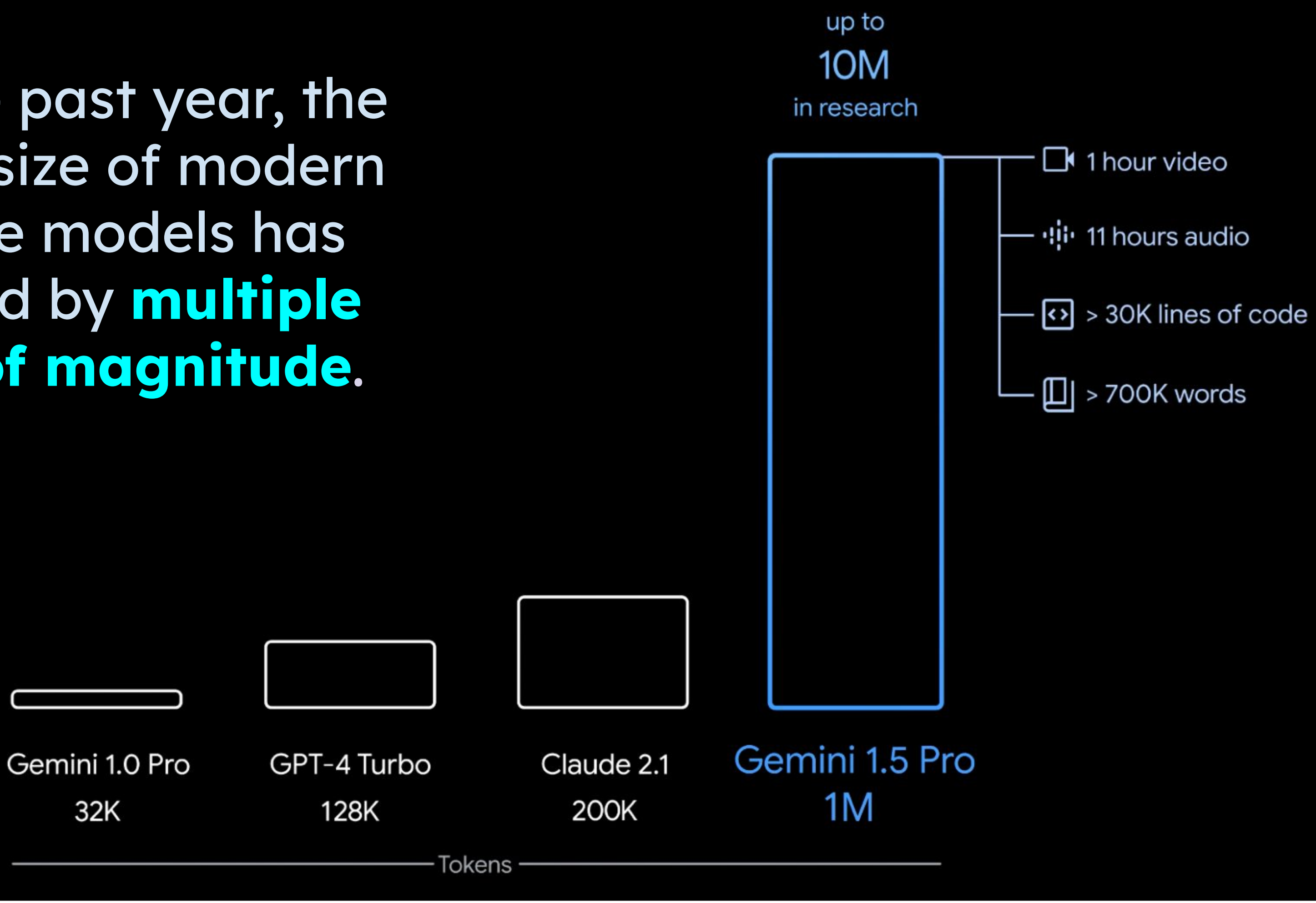
How can long-context language models improve RAG?

Mohit Iyyer
UMass Amherst

Over the past year, the context size of modern language models has increased by **multiple orders of magnitude**.



https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

What does this mean for RAG?

→ Can fit more/longer retrieved documents into the prompt, so retrievers don’t have to be perfect

→ Can potentially shift to cheap / fast retrievers as a result, e.g., BM25

But how well do LLMs *really* utilize these long contexts?

LLM context size

LLM reasoning abilities

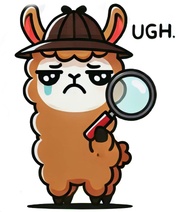
Evaluation of LLMs

Critical evaluation challenges with long input documents

→ Lack of reliable automatic metrics

→ Expensive and time-consuming to do human eval

→ Data contamination



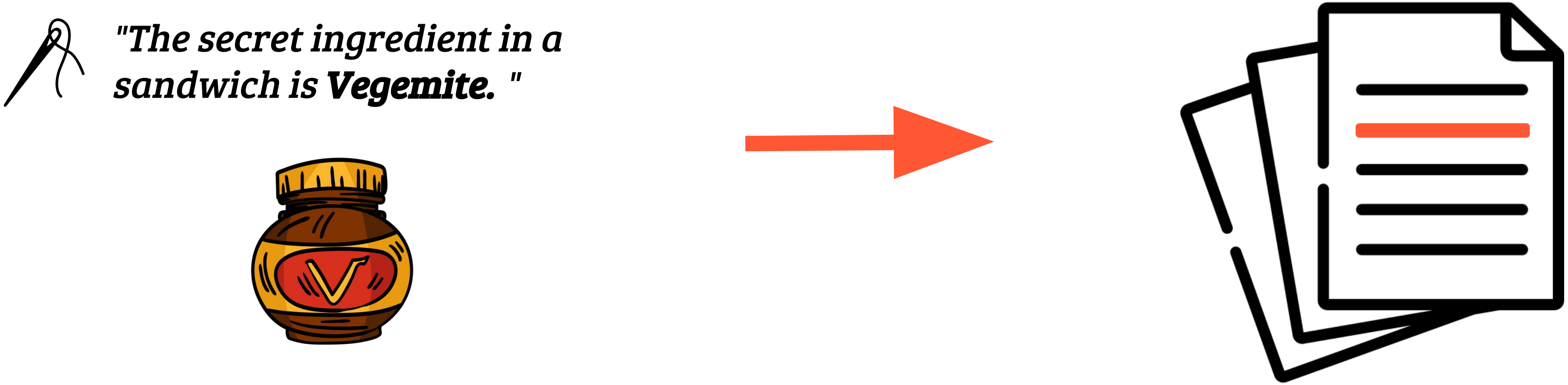
One way to get around these challenges is to create **synthetic** evaluations

“Needle in a Haystack” (NIAH)

"The secret ingredient in a sandwich is ***Vegemite.*** *"*

Kamradt (2023): https://github.com/gkamradt/LLMTest_NeedleInAHaystack

“Needle in a Haystack” (NIAH)



"The secret ingredient in a sandwich is ***Vegemite.*** *"*

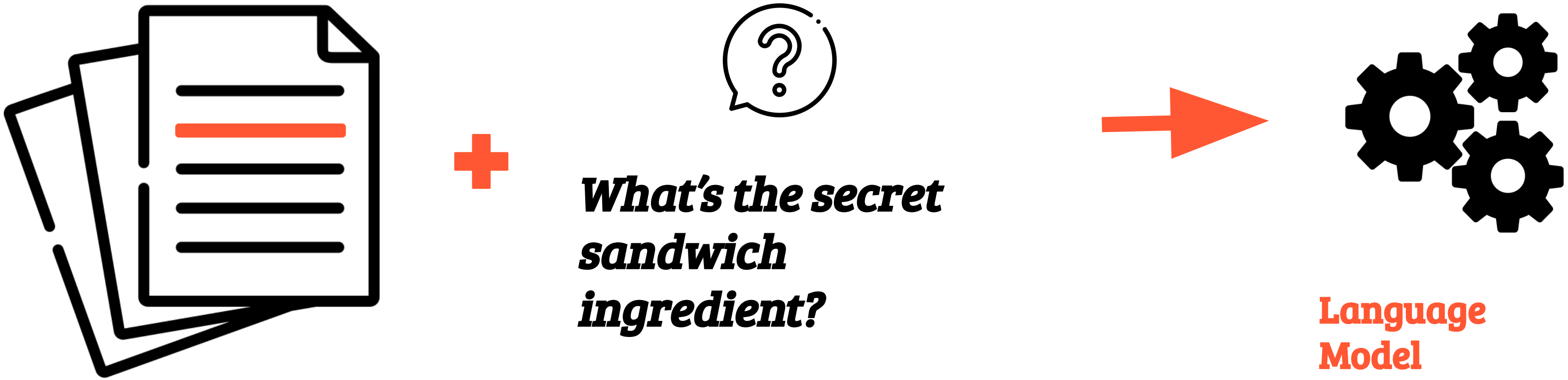
Kamradt (2023): https://github.com/gkamradt/LLMTest_NeedleInAHaystack

“Needle in a Haystack” (NIAH)

“Needle in a Haystack” (NIAH)

What’s the secret sandwich ingredient?

“Needle in a Haystack” (NIAH)

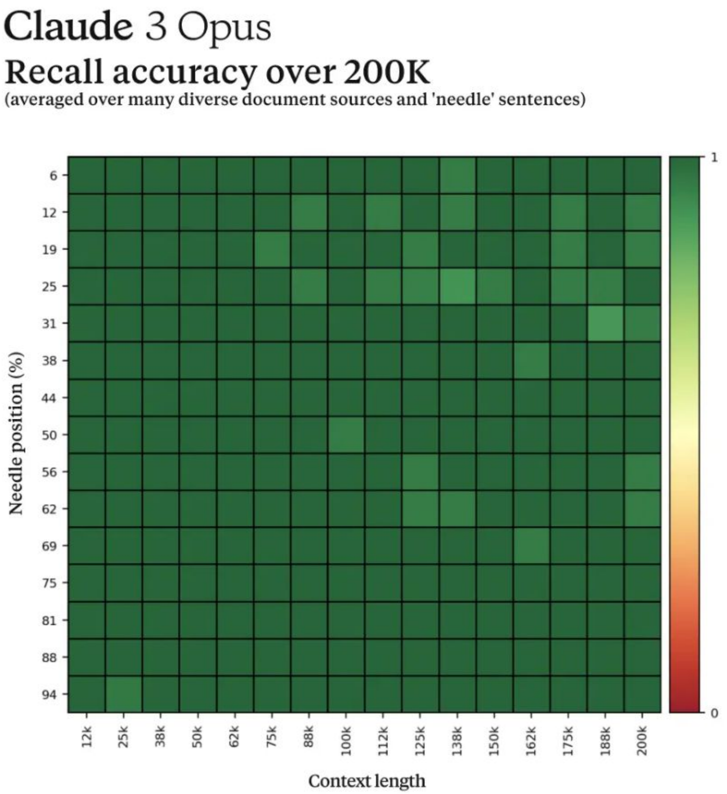


What’s the secret sandwich ingredient?

Language Model

Modern LLMs excel at NIAH!

"Claude 3 Opus not only achieved near-perfect recall, ***surpassing 99%*** *accuracy, but in some cases, it even* ***identified the limitations of the evaluation itself*** *by recognizing that the "needle" sentence* ***appeared to be artificially inserted into the original text by a human.*** *"*



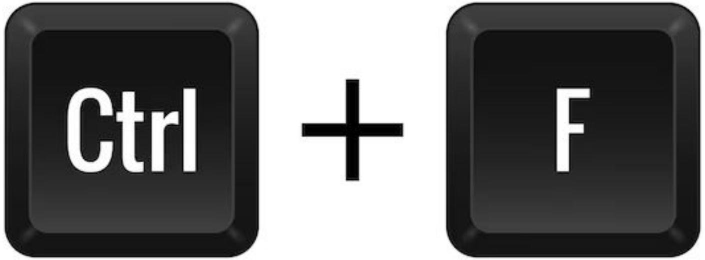
Claude 3 Opus

Recall accuracy over 200K

(averaged over many diverse document sources and 'needle' sentences)

https://www.anthropic.com/news/claude-3-family

Unfortunately, NIAH does not test complex retrieval, reasoning, or information synthesis.



IN THIS TALK:

- **Summarization** of recently published fictional books

- **Claim verification** over recently published fictional books

… okay, why “recently published fictional books”?

→ Using *recently published* books mitigates data contamination

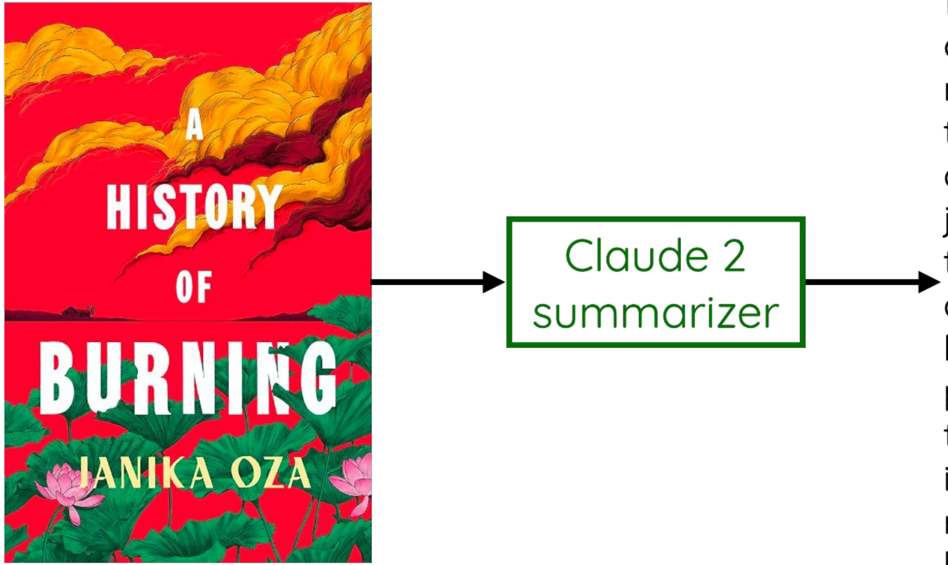
→ Using *fictional* books minimizes the LLM’s reliance on parametric knowledge

→ We choose books that annotators have *already* read for fun, thus minimizing annotation time / effort

FABLES: Evaluating factuality in book summarization (COLM 2024)

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, Mohit Iyyer.

LLMs can summarize long documents…

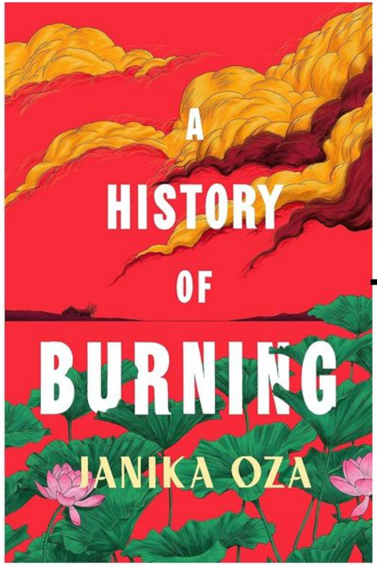


170K tokens

Claude 2 summarizer

In summary, the story spans 20 years and two families, exploring how broader political conflicts intertwine with and impact family relationships. In 1970s Uganda, rising political tensions strain the marriage of activists Latika and Arun as they clash over family versus justice. Two decades later in 1990s Canada, the family of Rajni and Vinod draw closer amidst violent unrest, losing their shop but keeping their bond intact. Despite different places and contexts, both families experience the complex dynamics between individual ideals, family ties, and turbulent political realities. Ultimately they share the resilient hope that as long as their family is together, they have arrived home

… but how faithful are these summaries?



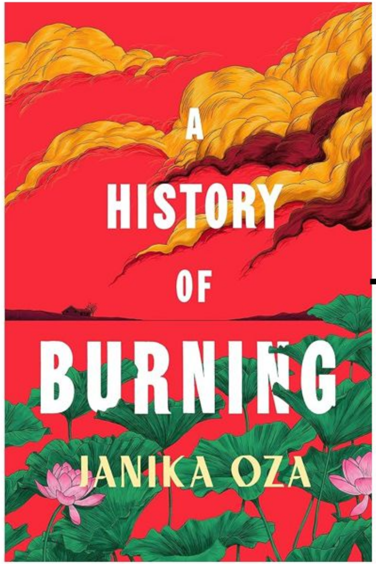
170K tokens

Claude 2 summarizer

104 years

In summary, the **story spans 20 years** and two families, exploring how broader political conflicts intertwine with and impact family relationships. In 1970s Uganda, rising political tensions strain the marriage of activists Latika and Arun as they clash over family versus justice. Two decades later in 1990s Canada, the family of Rajni and Vinod draw closer amidst violent unrest, losing their shop but keeping their bond intact. Despite different places and contexts, both families experience the complex dynamics between individual ideals, family ties, and turbulent political realities. Ultimately they share the resilient hope that as long as their family is together, they have arrived home

… but how faithful are these summaries?



170K tokens

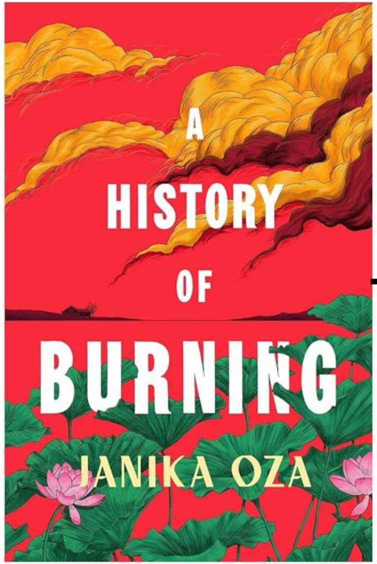
Claude 2 summarizer

One family, four generations ✗

✗ 104 years

In summary, the **story spans 20 years** and **two families**, exploring how broader political conflicts intertwine with and impact family relationships. In 1970s Uganda, rising political tensions strain the marriage of activists Latika and Arun as they clash over family versus justice. Two decades later in 1990s Canada, the family of Rajni and Vinod draw closer amidst violent unrest, losing their shop but keeping their bond intact. Despite different places and contexts, both families experience the complex dynamics between individual ideals, family ties, and turbulent political realities. Ultimately they share the resilient hope that as long as their family is together, they have arrived home

… but how faithful are these summaries?



170K tokens

Claude 2 summarizer

One family, four generations

104 years

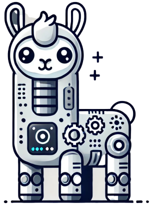
In summary, the **story spans 20 years** and **two families**, exploring how broader political conflicts intertwine with and impact family relationships. In 1970s Uganda, rising political tensions strain the marriage of activists Latika and Arun as they clash over family versus justice. Two decades later in 1990s Canada, **the family of Rajni and Vinod** draw closer amidst violent unrest, losing their shop but

Omits connection between the couples

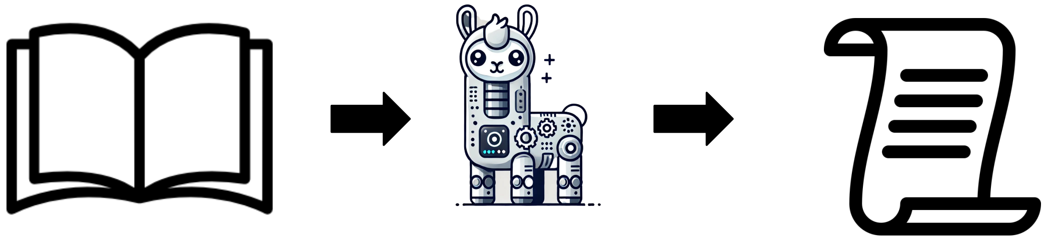
places and contexts, both families experience the complex dynamics between individual ideals, family ties, and turbulent political realities. Ultimately they share the resilient hope that as long as their family is together, they have arrived home

26 New Books

26 New Books



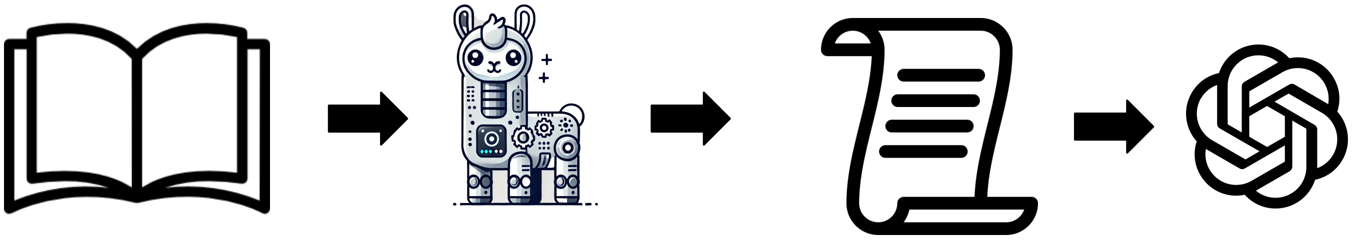
- 🔓Mixtral 8x7B
- 🔒GPT-3.5-Turbo
- 🔒GPT-4
- 🔒GPT-4-Turbo
- 🔒Claude-3-Opus



26 New Books

- 🔓 Mixtral 8x7B
- 🔒 GPT-3.5-Turbo
- 🔒 GPT-4
- 🔒 GPT-4-Turbo
- 🔒 Claude-3-Opus

Summary

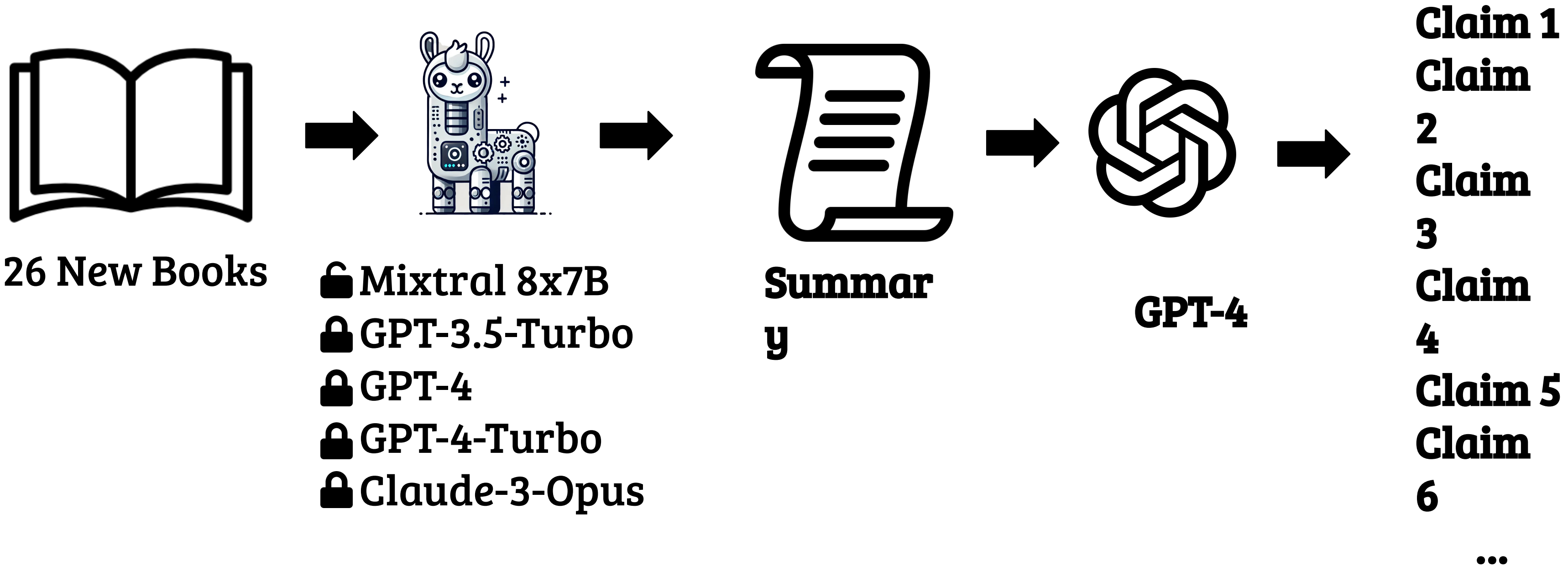


26 New Books

🔓 Mixtral 8x7B
🔒 GPT-3.5-Turbo
🔒 GPT-4
🔒 GPT-4-Turbo
🔒 Claude-3-Opus

Summary

GPT-4



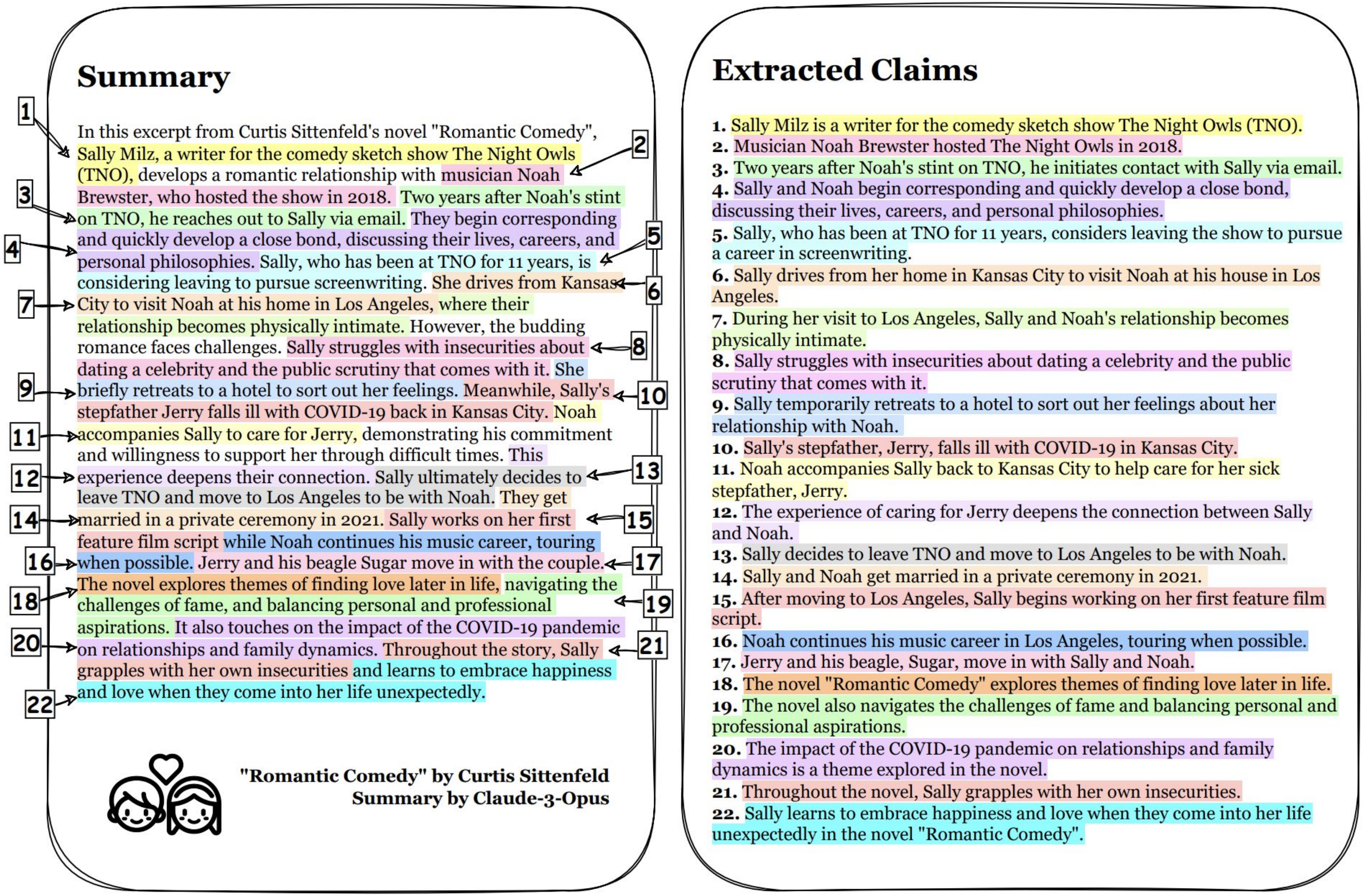
26 New Books

🔓 Mixtral 8x7B
🔒 GPT-3.5-Turbo
🔒 GPT-4
🔒 GPT-4-Turbo
🔒 Claude-3-Opus

Summary

GPT-4

Claim 1
Claim 2
Claim 3
Claim 4
Claim 5
Claim 6
...



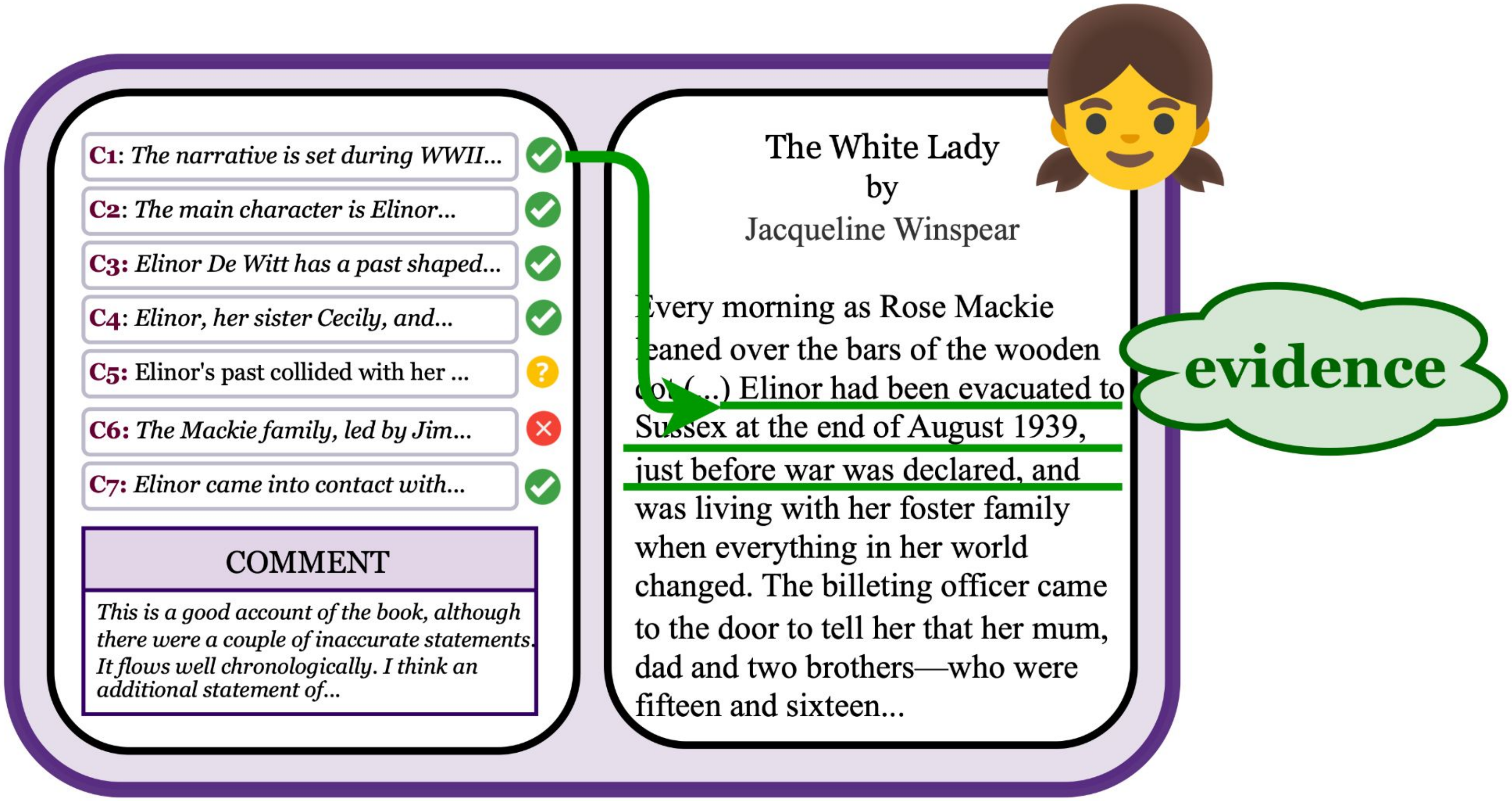
Summary

In this excerpt from Curtis Sittenfeld's novel "Romantic Comedy", Sally Milz, a writer for the comedy sketch show The Night Owls (TNO), develops a romantic relationship with musician Noah Brewster, who hosted the show in 2018. Two years after Noah's stint on TNO, he reaches out to Sally via email. They begin corresponding and quickly develop a close bond, discussing their lives, careers, and personal philosophies. Sally, who has been at TNO for 11 years, is considering leaving to pursue screenwriting. She drives from Kansas City to visit Noah at his home in Los Angeles, where their relationship becomes physically intimate. However, the budding romance faces challenges. Sally struggles with insecurities about dating a celebrity and the public scrutiny that comes with it. She briefly retreats to a hotel to sort out her feelings. Meanwhile, Sally's stepfather Jerry falls ill with COVID-19 back in Kansas City. Noah accompanies Sally to care for Jerry, demonstrating his commitment and willingness to support her through difficult times. This experience deepens their connection. Sally ultimately decides to leave TNO and move to Los Angeles to be with Noah. They get married in a private ceremony in 2021. Sally works on her first feature film script while Noah continues his music career, touring when possible. Jerry and his beagle Sugar move in with the couple. The novel explores themes of finding love later in life, navigating the challenges of fame, and balancing personal and professional aspirations. It also touches on the impact of the COVID-19 pandemic on relationships and family dynamics. Throughout the story, Sally grapples with her own insecurities and learns to embrace happiness and love when they come into her life unexpectedly.

"Romantic Comedy" by Curtis Sittenfeld
Summary by Claude-3-Opus

Extracted Claims

1. Sally Milz is a writer for the comedy sketch show The Night Owls (TNO).
2. Musician Noah Brewster hosted The Night Owls in 2018.
3. Two years after Noah's stint on TNO, he initiates contact with Sally via email.
4. Sally and Noah begin corresponding and quickly develop a close bond, discussing their lives, careers, and personal philosophies.
5. Sally, who has been at TNO for 11 years, considers leaving the show to pursue a career in screenwriting.
6. Sally drives from her home in Kansas City to visit Noah at his house in Los Angeles.
7. During her visit to Los Angeles, Sally and Noah's relationship becomes physically intimate.
8. Sally struggles with insecurities about dating a celebrity and the public scrutiny that comes with it.
9. Sally temporarily retreats to a hotel to sort out her feelings about her relationship with Noah.
10. Sally's stepfather, Jerry, falls ill with COVID-19 in Kansas City.
11. Noah accompanies Sally back to Kansas City to help care for her sick stepfather, Jerry.
12. The experience of caring for Jerry deepens the connection between Sally and Noah.
13. Sally decides to leave TNO and move to Los Angeles to be with Noah.
14. Sally and Noah get married in a private ceremony in 2021.
15. After moving to Los Angeles, Sally begins working on her first feature film script.
16. Noah continues his music career in Los Angeles, touring when possible.
17. Jerry and his beagle, Sugar, move in with Sally and Noah.
18. The novel "Romantic Comedy" explores themes of finding love later in life.
19. The novel also navigates the challenges of fame and balancing personal and professional aspirations.
20. The impact of the COVID-19 pandemic on relationships and family dynamics is a theme explored in the novel.
21. Throughout the novel, Sally grapples with her own insecurities.
22. Sally learns to embrace happiness and love when they come into her life unexpectedly in the novel "Romantic Comedy".



C1: *The narrative is set during WWII...* ✓

C2: *The main character is Elinor...* ✓

C3: *Elinor De Witt has a past shaped...* ✓

C4: *Elinor, her sister Cecily, and...* ✓

C5: Elinor's past collided with her ... ?

C6: *The Mackie family, led by Jim...* ✗

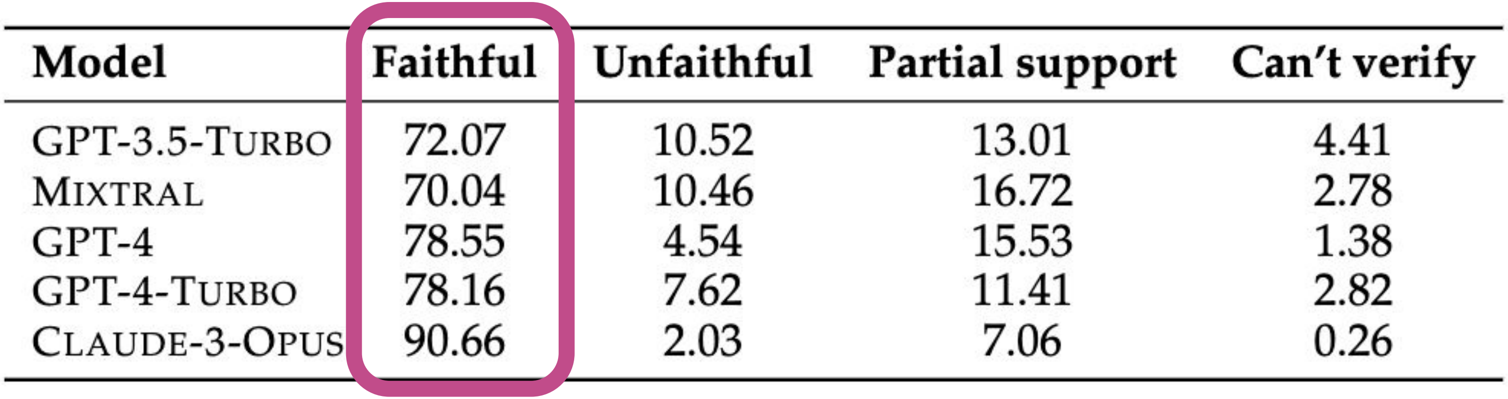
C7: *Elinor came into contact with...* ✓

COMMENT

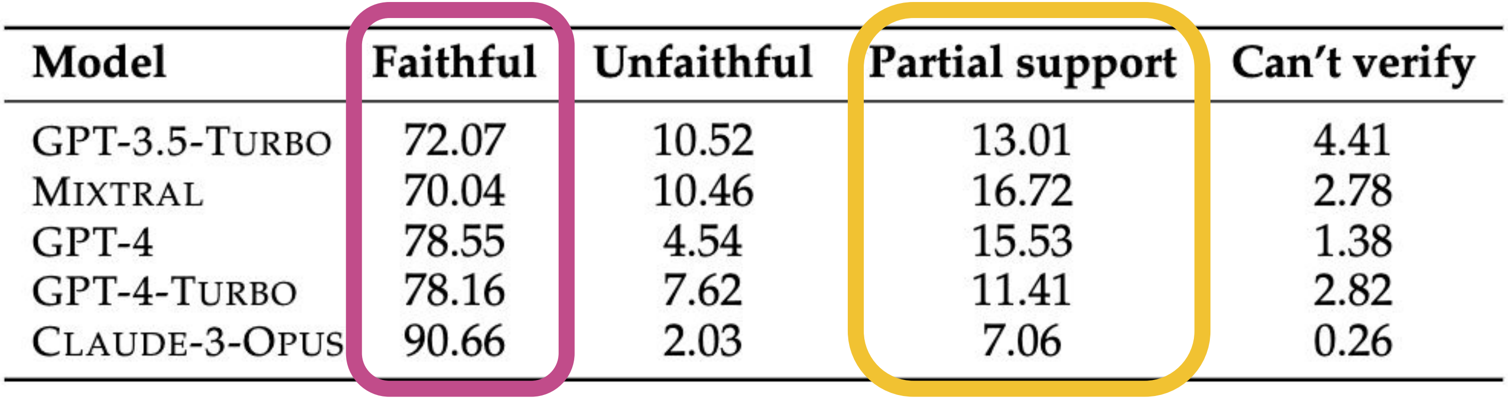
This is a good account of the book, although there were a couple of inaccurate statements. It flows well chronologically. I think an additional statement of...

The White Lady
by
Jacqueline Winspear

Every morning as Rose Mackie leaned over the bars of the wooden cot (...) Elinor had been evacuated to Sussex at the end of August 1939, just before war was declared, and was living with her foster family when everything in her world changed. The billeting officer came to the door to tell her that her mum, dad and two brothers—who were fifteen and sixteen...



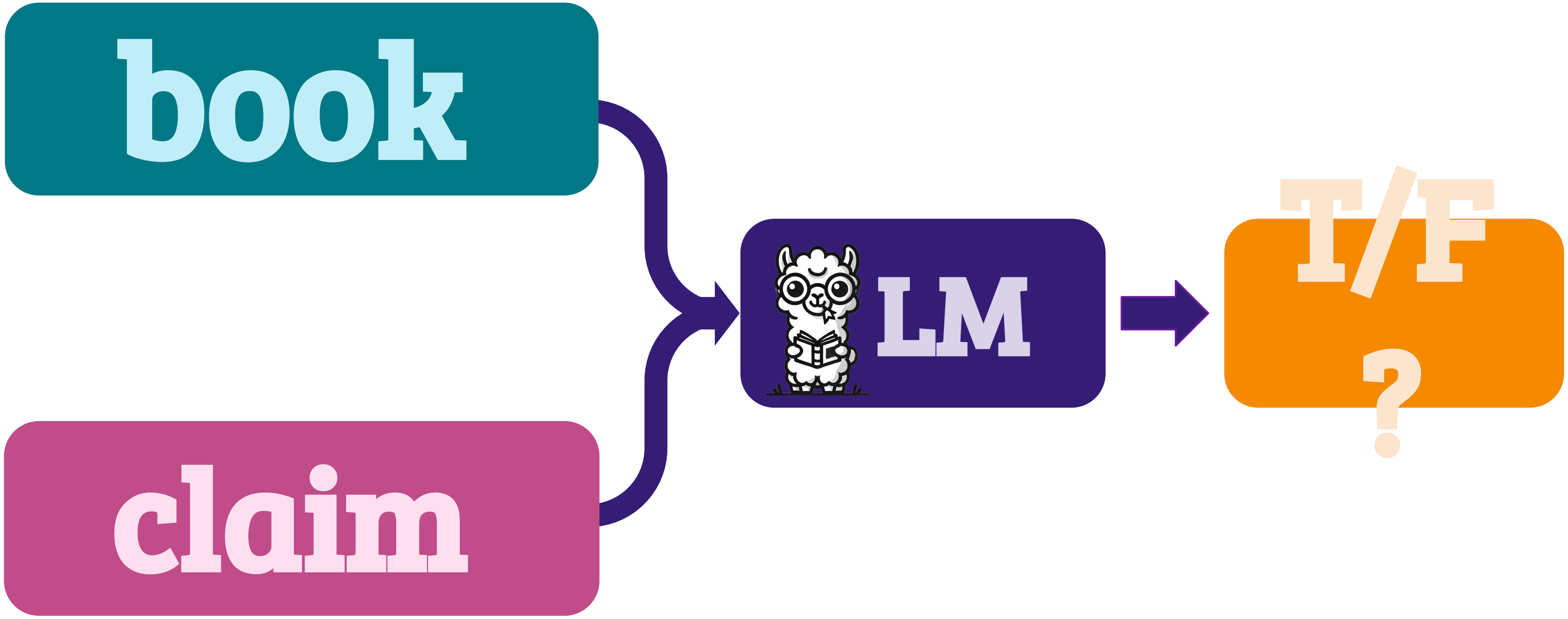
Model	Faithful	Unfaithful	Partial support	Can’t verify
GPT-3.5-Turbo	72.07	10.52	13.01	4.41
Mixtral	70.04	10.46	16.72	2.78
GPT-4	78.55	4.54	15.53	1.38
GPT-4-Turbo	78.16	7.62	11.41	2.82
Claude-3-Opus	90.66	2.03	7.06	0.26



Model	Faithful	Unfaithful	Partial support	Can't verify
GPT-3.5-Turbo	72.07	10.52	13.01	4.41
Mixtral	70.04	10.46	16.72	2.78
GPT-4	78.55	4.54	15.53	1.38
GPT-4-Turbo	78.16	7.62	11.41	2.82
Claude-3-Opus	90.66	2.03	7.06	0.26

Label	Freq	Example claim	Reason for rejection
		Claim Type	
State	38.6	*Roman Kitt is under pressure from his father to join the family business.*	Roman is not under pressure, his father bribes people so he gets his dream job.
Event	31.5	*Patricia Liu, Athena's mother, discovers that June has sold Athena's manuscript and confronts her.*	Patricia never confronts June.
Cause/effect	11.2	*Lilly's abusive ex-boyfriend, Alan Bushy, becomes a suspect due to the meticulous nature of the murders.*	He becomes a suspect because he was abusive to Lilly.

Can we use LLMs to evaluate the factuality of generated claims?



book

claim

LM

T/F

?

Lilith has a chance encounter with a vampire named Vale.

Lilith has a chance encounter with a vampire named Vale.

FALSE: *It wasn't a chance encounter, Lilith knew whom she was going to meet.*

Lilith has a chance encounter with a vampire named Vale.

FALSE: *It wasn't a chance encounter, Lilith knew whom she was going to meet.*

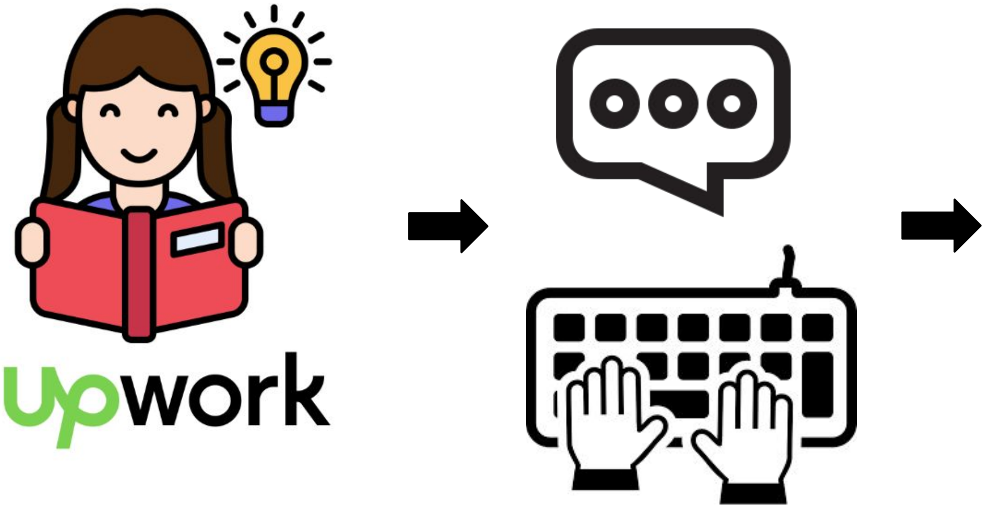
True

True

NoCha: A Novel Challenge for long-context LLMs

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, Mohit Iyyer.

One Thousand and One Pairs: A"novel"challenge for long-context language models.



TRUE

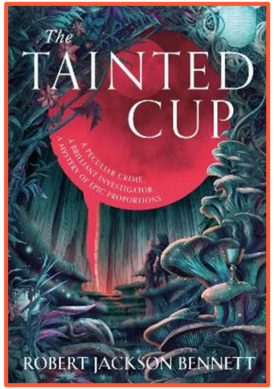
Despite her skills as an Apoth , Nusis is unable to reverse engineer the type of portal opened by the reagents key found in Rona's wooden chest.

FALSE

By using her skills as an Apoth , Nusis is able to reverse engineer the type of portal opened by the reagents key found in Rona's wooden chest.

!

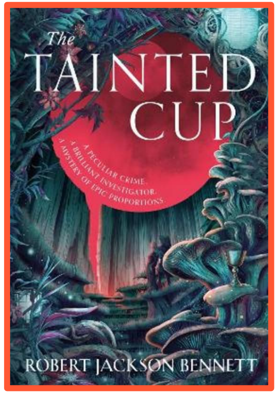
The reagents key is in fact not a key at all but the cure for dappleglass poisoning, which explains why Nusis is unable to figure out what type of portal it opens.



+ TRUE

Despite her skills as an Apoth , Nusis is unable to reverse engineer the type of portal opened by the reagents key found in Rona's wooden chest..

→ → TRUE



+ FALSE

By using her skills as an Apoth , Nusis is able to reverse engineer the type of portal opened by the reagents key found in Rona's wooden chest.

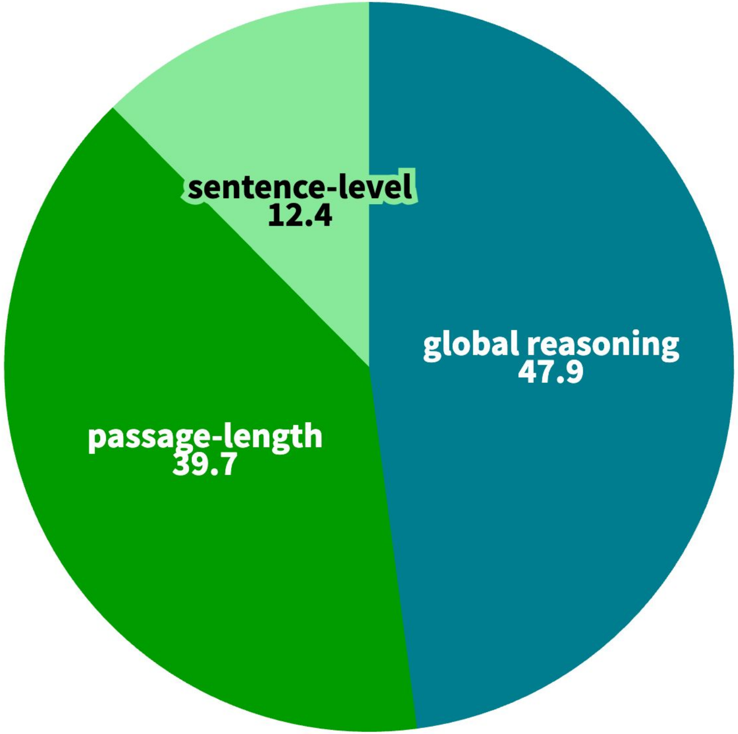
→ → FALSE

NoCha dataset

→ **67 books: 30 books** published in 2024, **33 books** published in 2023, and **4 classics**

→ **1,001 claim pairs** in total, ~10-15 pairs per book

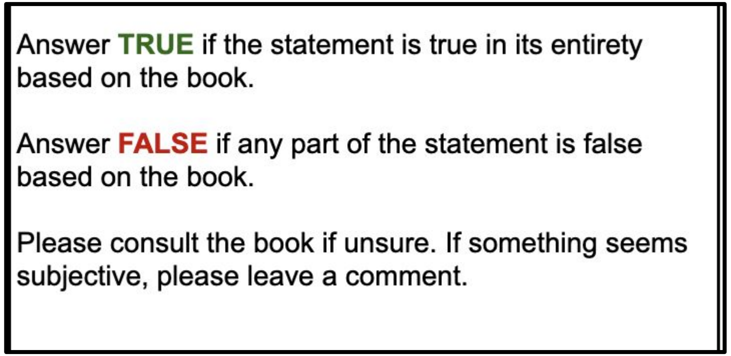
Most claims in NoCha require **global reasoning**



Human performance on NoCha is extremely high!

→ Verified claims for **6 books** for which we could get two readers

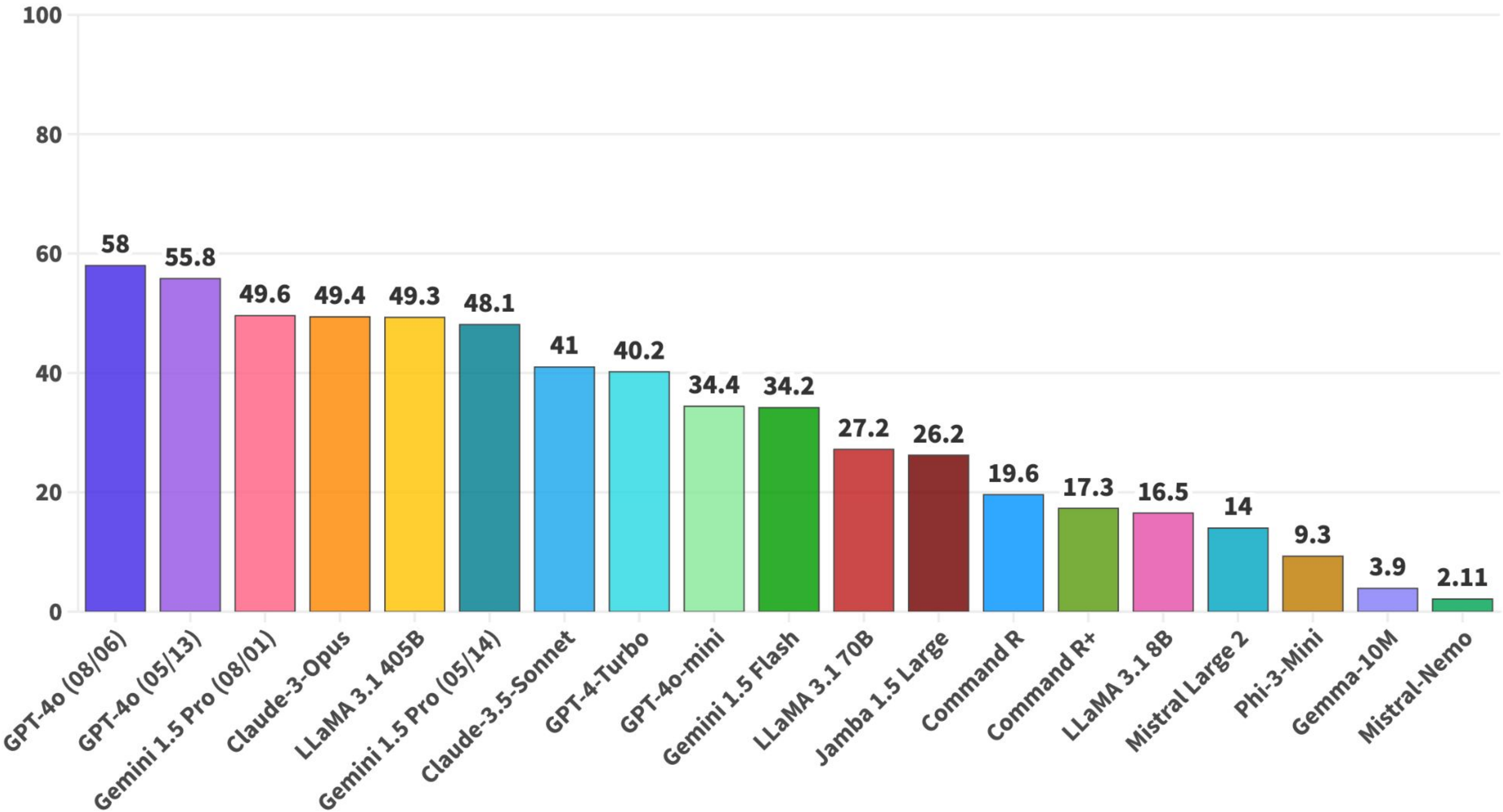
→ Overall, annotators agreed on **97.4%** of claim pairs!

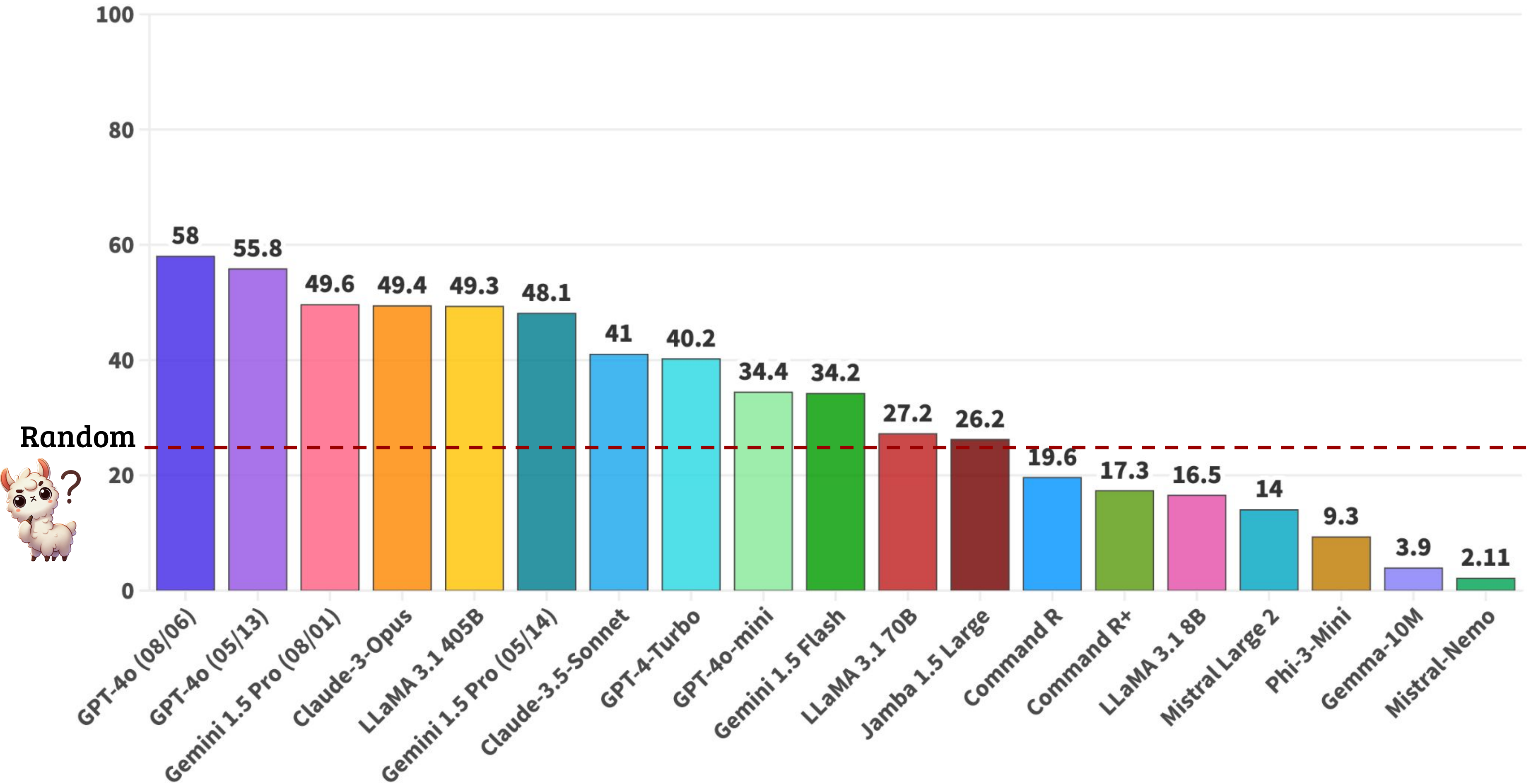


Answer **TRUE** if the statement is true in its entirety based on the book.

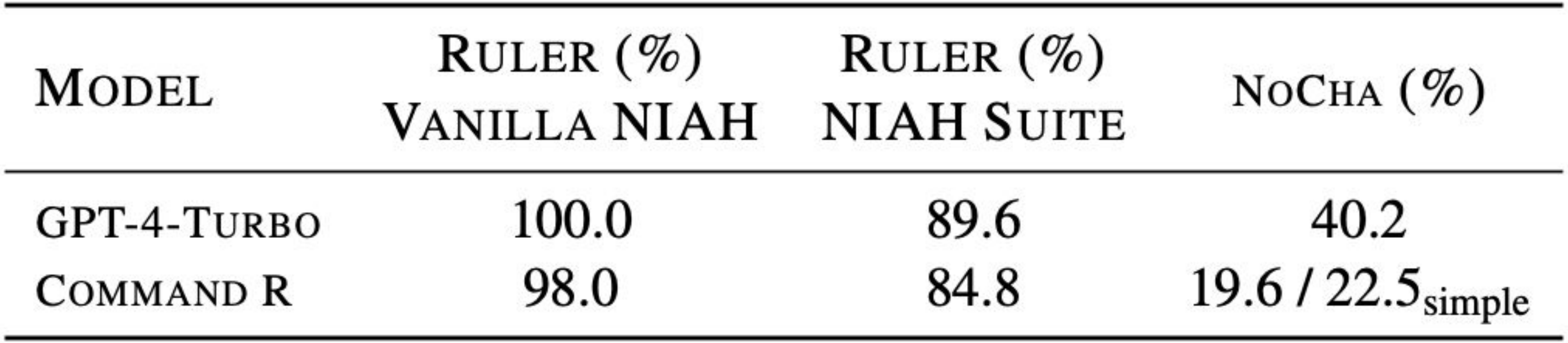
Answer **FALSE** if any part of the statement is false based on the book.

Please consult the book if unsure. If something seems subjective, please leave a comment.


100
80
60
40
20
0
58
55.8
49.6
49.4
49.3
48.1
41
40.2
34.4
34.2
27.2
26.2
19.6
17.3
16.5
14
9.3
3.9
2.11
GPT-4o (08/06)
GPT-4o (05/13)
Gemini 1.5 Pro (08/01)
Claude-3-Opus
LLaMA 3.1 405B
Gemini 1.5 Pro (05/14)
Claude-3.5-Sonnet
GPT-4-Turbo
GPT-4o-mini
Gemini 1.5 Flash
LLaMA 3.1 70B
Jamba 1.5 Large
Command R
Command R+
LLaMA 3.1 8B
Mistral Large 2
Phi-3-Mini
Gemma-10M
Mistral-Nemo


100
80
60
40
20
0
Random
58
55.8
49.6
49.4
49.3
48.1
41
40.2
34.4
34.2
27.2
26.2
19.6
17.3
16.5
14
9.3
3.9
2.11
GPT-4o (08/06)
GPT-4o (05/13)
Gemini 1.5 Pro (08/01)
Claude-3-Opus
LLaMA 3.1 405B
Gemini 1.5 Pro (05/14)
Claude-3.5-Sonnet
GPT-4-Turbo
GPT-4o-mini
Gemini 1.5 Flash
LLaMA 3.1 70B
Jamba 1.5 Large
Command R
Command R+
LLaMA 3.1 8B
Mistral Large 2
Phi-3-Mini
Gemma-10M
Mistral-Nemo

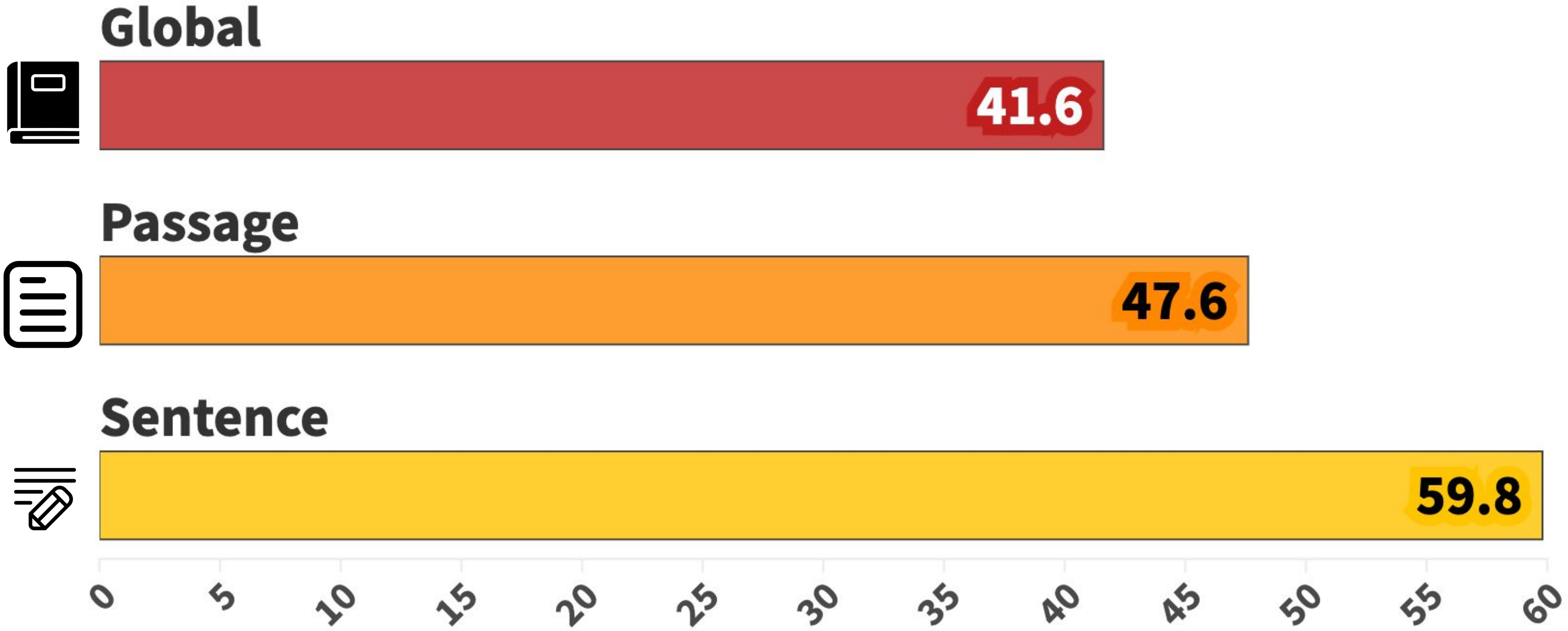
How does NoCha compare to NIAH?



Model	Ruler (%) Vanilla NIAH	Ruler (%) NIAH Suite	NoCha (%)
GPT-4-Turbo	100.0	89.6	40.2
Command R	98.0	84.8	19.6 / 22.5$_{\text{simple}}$

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models?

NoCha claims requiring **global reasoning** are hard to verify

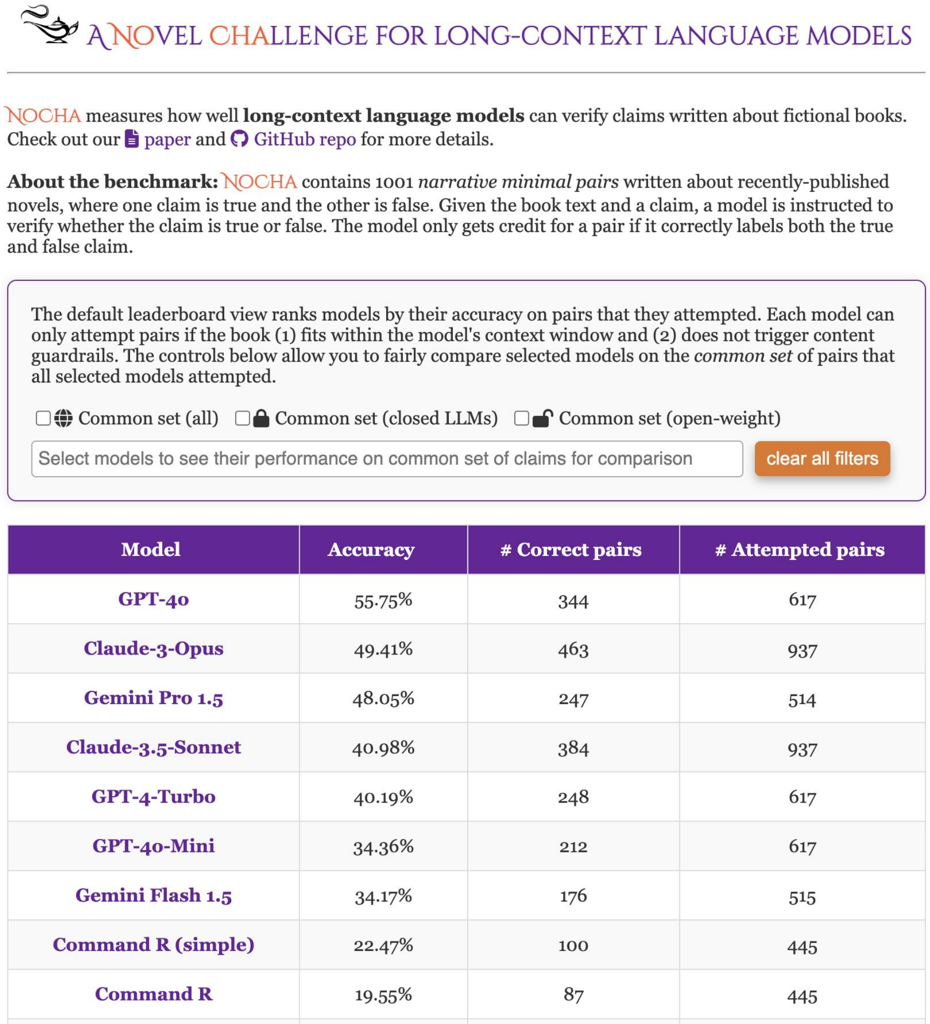


What does this mean for RAG?

→ Long context LLMs are not yet able to fully understand and process their inputs

→ Currently, building stronger retrievers and feeding a short list of documents to an LLM is likely a better option

→ But how long will this state of affairs last?



A NOVEL CHALLENGE FOR LONG-CONTEXT LANGUAGE MODELS

NOCHA measures how well **long-context language models** can verify claims written about fictional books. Check out our paper and GitHub repo for more details.

About the benchmark: NOCHA contains 1001 *narrative minimal pairs* written about recently-published novels, where one claim is true and the other is false. Given the book text and a claim, a model is instructed to verify whether the claim is true or false. The model only gets credit for a pair if it correctly labels both the true and false claim.

The default leaderboard view ranks models by their accuracy on pairs that they attempted. Each model can only attempt pairs if the book (1) fits within the model's context window and (2) does not trigger content guardrails. The controls below allow you to fairly compare selected models on the *common set* of pairs that all selected models attempted.

□ Common set (all) □ Common set (closed LLMs) □ Common set (open-weight)

Select models to see their performance on common set of claims for comparison

Model	Accuracy	# Correct pairs	# Attempted pairs
GPT-4o	55.75%	344	617
Claude-3-Opus	49.41%	463	937
Gemini Pro 1.5	48.05%	247	514
Claude-3.5-Sonnet	40.98%	384	937
GPT-4-Turbo	40.19%	248	617
GPT-4o-Mini	34.36%	212	617
Gemini Flash 1.5	34.17%	176	515
Command R (simple)	22.47%	100	445
Command R	19.55%	87	445

Thanks!!!
Questions?

https://novelchallenge.github.io