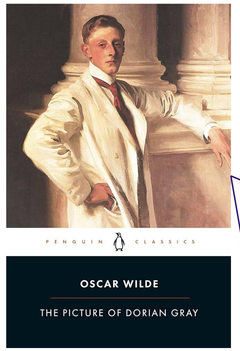
✂️ CLIPPER: Compression enables long-context synthetic data generation

Chau Minh Pham 💇 • Yapei Chang 💇 • Mohit Iyyer 💇💈 (💇University of Maryland, College Park • 💈University of Massachusetts Amherst)

TLDR: A **compression-based** approach for generating **synthetic data** for ***narrative claim verification***

Task: A model needs to decide if a claim about a book is True or False and explain why.

Stage 1: Compression



Book text
74K tokens

Generating claims directly from books raises error rates from 17% to 73%, with spikes in misattributed (29%) & duplicated claims (17%)!

Chapter 1

📝 *Synopsis*: Lord Henry Wotton visits artist Basil Hallward's studio...

🎭 *Events*: Lord Henry Wotton observes a portrait of a young man...

👤 *Characters*:
- Lord Henry Wotton: Basil's friend, cynical and witty
- Basil Hallward: the artist

Chapter outline
7K tokens

A handsome young man, Dorian Gray, wishes his portrait would age instead of him. His wish comes true: his portrait grows monstrous, reflecting his inner depravity, while he remains youthful...

Book summary
1K tokens

Stage 2: Synthetic claim generation

✅ Dorian Gray's relationship with Sibyl Vane progresses from his confession of love to her **poor** performance as Juliet, leading to his **cruel rejection**.

❌ Dorian Gray's relationship with Sibyl Vane progresses from his confession of love to her **excellent** performance as Juliet, leading to their **marriage**.

Book-level claim pairs

<*explanation*>
- Chapter 4: Dorian falls for Sibyl Vane.
- Chapter 7: Her poor Juliet performance disappoints Dorian; he cruelly rejects her.
- Chapter 8: Lord Henry tells Dorian of Sibyl's suicide.
- Relationship: Events in Dorian and Sibyl's relationship.

</*explanation*>

<*answer*>True</*answer*>

Chain-of-thought explanation

Fine-tuning on CoTs results in more informative & grounded explanations!

✅ Lord Henry **praises** Basil Hallward's portrait of Dorian Gray and suggests **exhibiting** it, but Basil refuses, citing **personal reasons.**

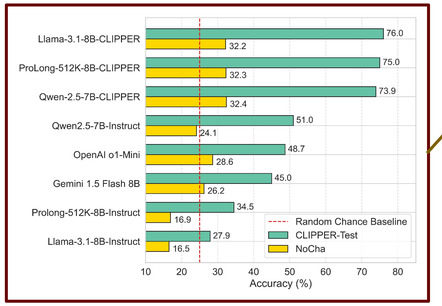
❌ Lord Henry **criticizes** Basil Hallward's portrait of Dorian Gray and suggests **destroying** it, but Basil refuses, citing its **monetary value.**

Chapter-level claim pairs

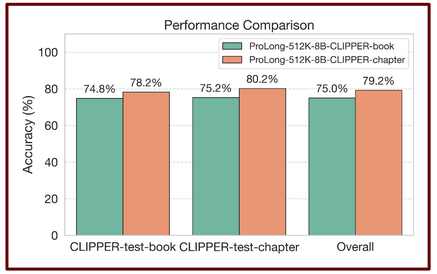
➤ CLIPPER models improve on reasoning tasks

Models	CLIPPER-test	NoCha	Narrative QA	MuSR	∞Bench QA
Qwen2.5-7B-Instruct	51.0%	24.1%	40.3%	41.2%	35.3%
Llama-3.1-8B-Instruct	27.9%	16.5%	47.7%	40.3%	**47.8%**
ProLong-512K-8B-Instruct	34.5%	16.9%	44.0%	42.3%	42.6%
✂️ Qwen2.5-7B-CLIPPER	73.9%	**32.4%**	46.0%	45.2%	42.3%
✂️ Llama-3.1-8B-CLIPPER	76.0%	32.2%	**49.0%**	43.6%	46.5%
✂️ ProLong-512K-8B-CLIPPER	75.0%	32.3%	**49.0%**	44.5%	38.5%
✂️ ProLong-512K-8B-WritingPrompts	63.0%	24.1%	31.0%	45.2%	35.8%

Fine-tuning on short-context claim data is less helpful for long-context narrative reasoning tasks!



CLIPPER models are state-of-the-art sub-10B models on NoCha!



Small models struggle with book-level reasoning...

Ongoing work

Using GRPO to improve long-context models on narrative claim verification tasks