Understanding Synthetic Context Extension via Retrieval Heads

Xinyu Zhao, Fangcong Yin, Greg Durrett

TAUR LAB @UT Austin

ICML International Conference On Machine Learning

How does training on synthetic data improve long-context models on real data?

Synthetic Data

[Doc 1] . . . ***WPFN is the place of death of ABCD****. . .*

. . .

Q: *What currency is used where ABCD died?*

A: *WXZY*

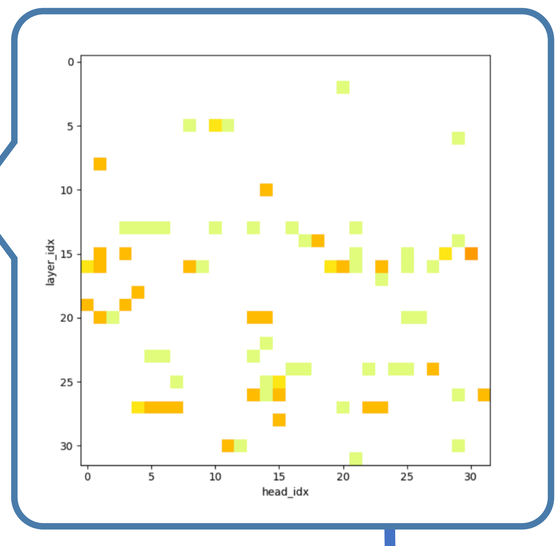
Real Data

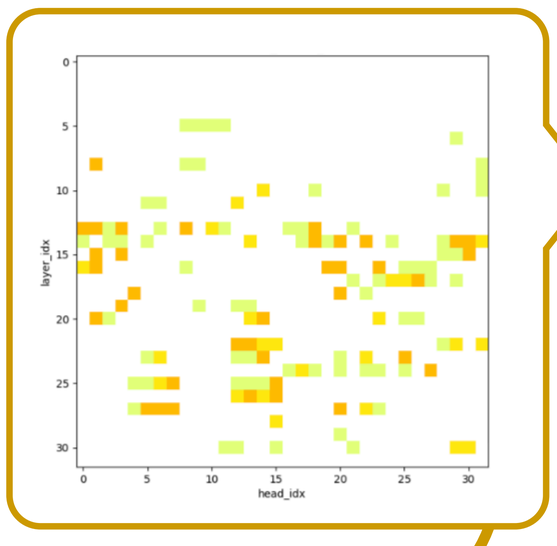
[Doc 1] Billy Giles was an Ulster Volunteer Force volunteer who later became active in politics. . . ***Billy Giles died on 25 September 1998 (aged 41) in Belfast, Northern Ireland, the United Kingdom****. . .*

. . . [Wikipedia Passages] . . .

Q: *What currency is used where Billy Giles died?*

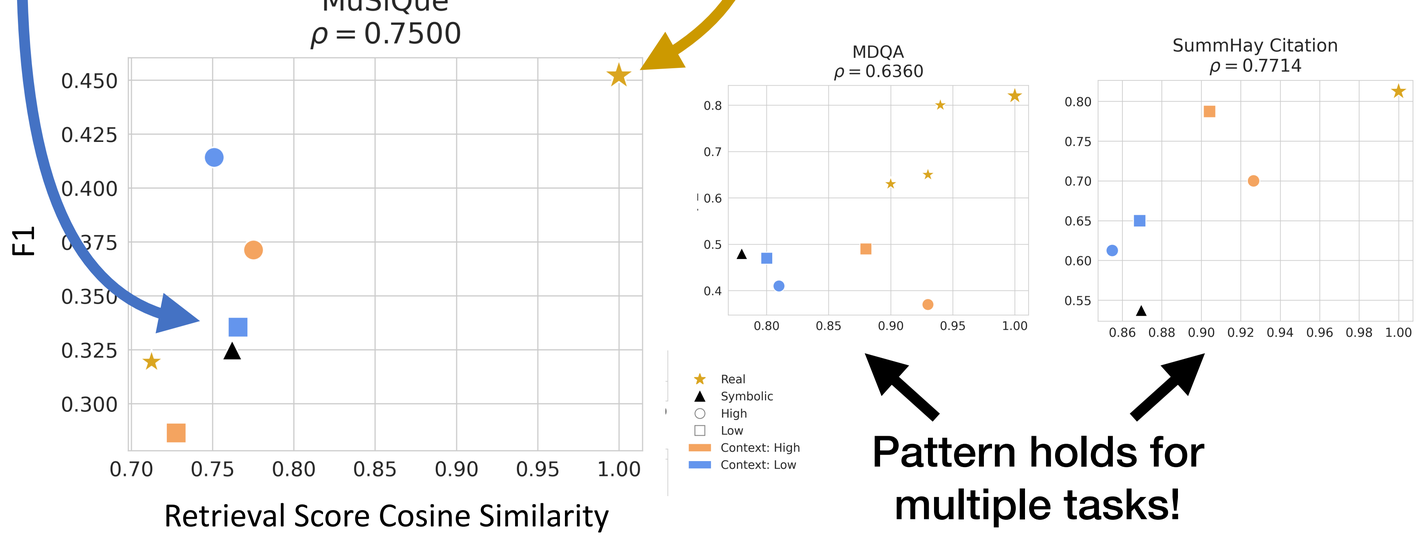
A: *Pound Sterling*

SFT → **Synthetic model** → 

SFT → **Real model** → 

Identify *retrieval heads* (attention heads which extract the answer)

Finding: Synthetic data that uses retrieval heads similar to real data yields better downstream performance!



Synthetic Dataset Construction

We vary both the **"needle" concept** and the **"haystack" context** realism:

Real Data

[Doc 1] *Billy Giles was an Ulster Volunteer Force volunteer who later became active in politics…* ***Billy Giles died on 25 September 1998 (aged 41) in Belfast, Northern Ireland, the United Kingdom****. Giles is commemorated, along with other prominent Loyalist paramilitaries…*

Concept (Needle) ← ***Billy Giles died on 25 September 1998 (aged 41) in Belfast, Northern Ireland, the United Kingdom***

Context (Haystack) ← *[Doc 1] Billy Giles was an Ulster Volunteer Force volunteer who later became active in politics… Giles is commemorated, along with other prominent Loyalist paramilitaries…*

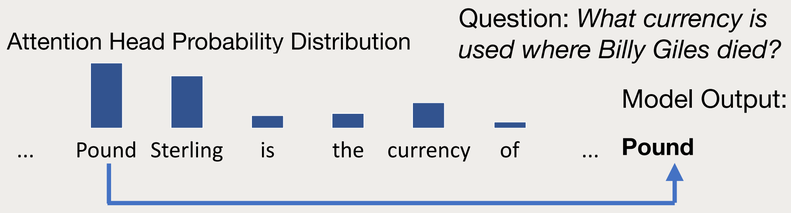
Synthetic Data

	Context	Concept
High Realism	*As a volunteer of Ulster Volunteer Force, Billy Giles later participated in political activities…*	*At the age of 41, Billy Giles died in Belfast, Northern Ireland, UK on 25 September 1998.*
Low Realism	*The grass is green. The sky is blue. The sun is yellow… The grass is green. The sky is blue…*	*WFPN is the place of death of ABCD.*
Symbolic	KVTJ { …, 'UQCA': 'SXVI', 'ERQG': 'FQDR', 'TZAM': 'XYTH', … }	

Our Hypothesis: Effective synthetic data trains the required attention heads for the given task

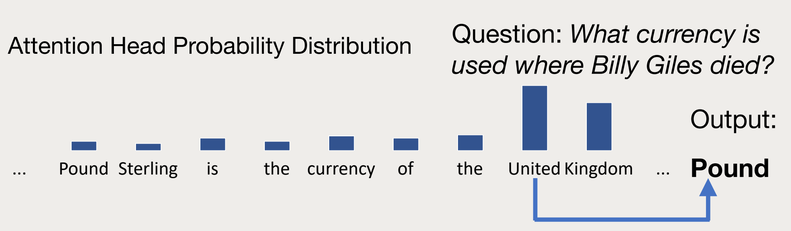
Retrieval Heads

An attention head is a retrieval head if it places highest attention probability on the answer in context

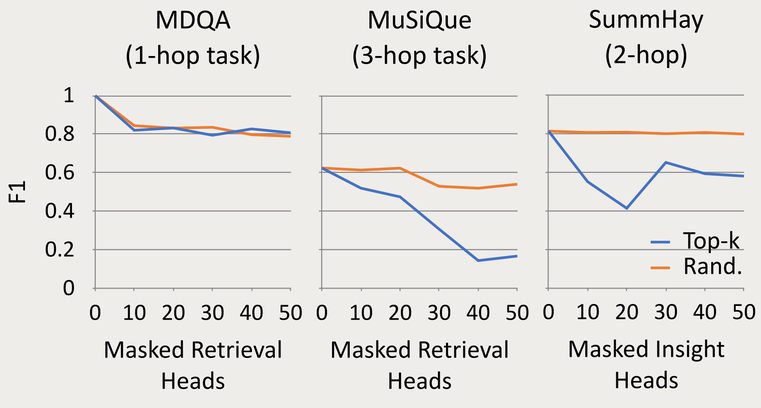


"Insight" Heads

Attention heads which place highest attention probability on intermediate reasoning "insights" in

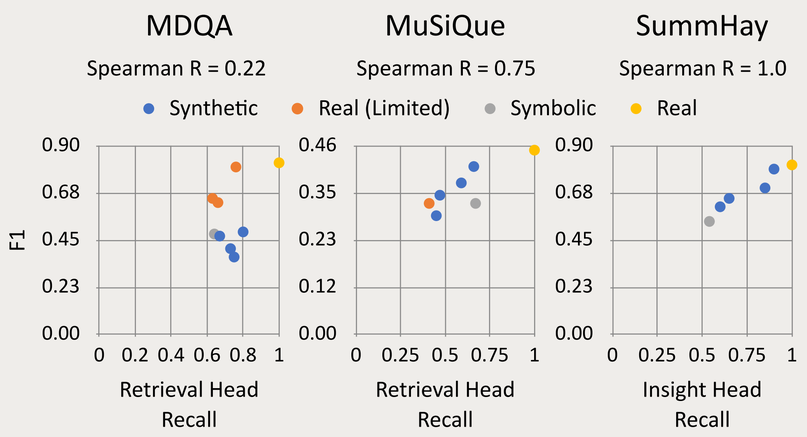


We study three long-context reasoning and retrieval tasks:

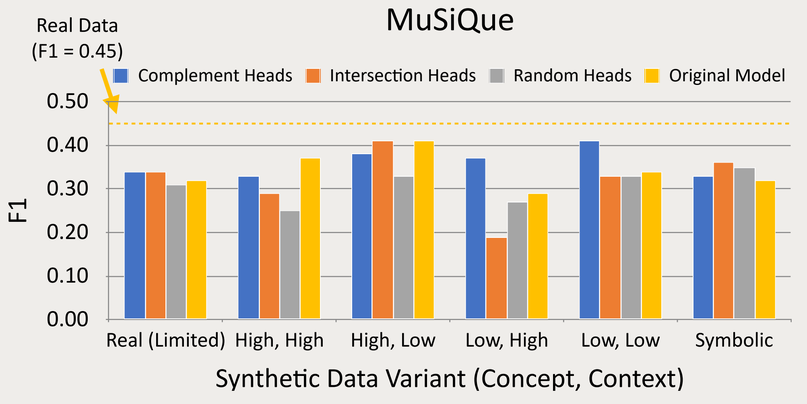


Even though these tasks vary in style and complexity, both retrieval and insight heads are required for task performance.

Finding: Retrieval Head Recall is also correlated with performance on the real dataset.



Finding: Synthetic data retrieval heads **are not as effective** as real data retrieval heads.



Findings are similar with both Llama-3-8B Instruct (shown on this poster) and Mistral-7b-Instruct-v0.1!

arXiv