THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Are Large Language Models Good Data Preprocessors?

ACM WWW'25 Workshop

Presenter: Elyas Meguellati

Co-authors:
Nardiena Pratama, Shazia Sadiq, Gianluca Demartini

Motivation and Problem

- High-quality text data is crucial for context dependent tasks
- Image captioning models (BLIP, GIT) often produce noisy captions
- Rule-based cleaning struggles with diverse errors

Question:

Can LLMs reliably clean and improve noisy text?

Scope

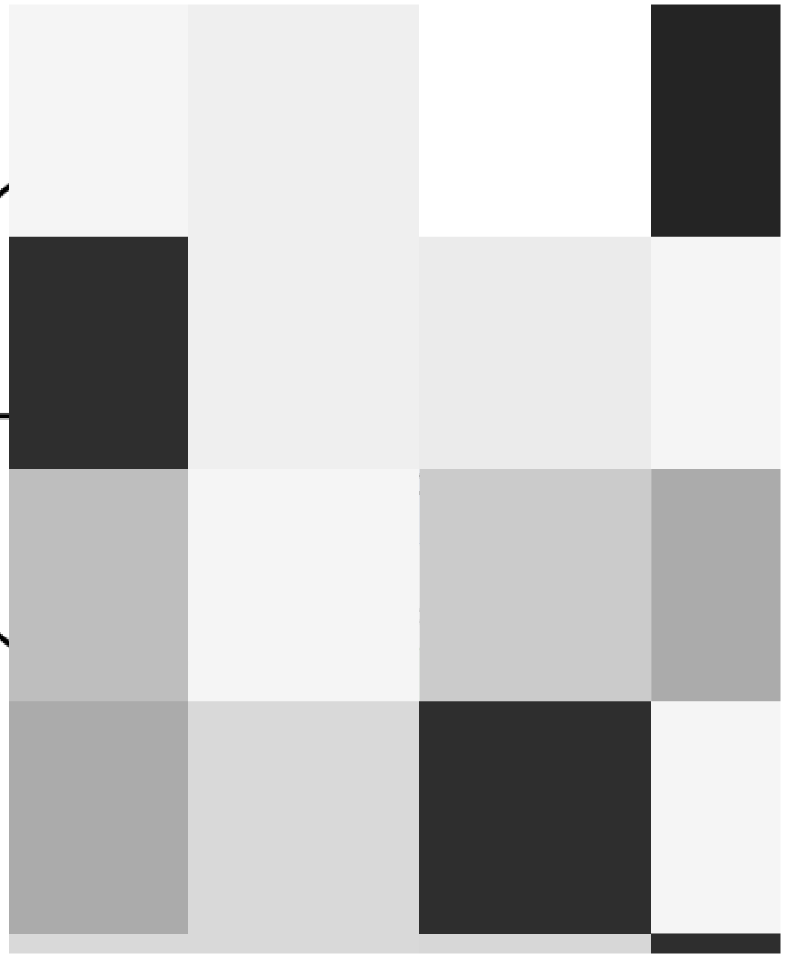
- Aim: Investigating LLMs effectiveness in cleaning text from image captioning models
- Dataset: Multi-label persuasion in memes (SemEval 2024 Task)
- Metric: Heirarchical F1 (order matters)
- ~~LLMs: LLaMA 3.1-70B, GPT-4 Turbo, Sonnet 3.5 v2~~

Blip: a horse with its mouth open

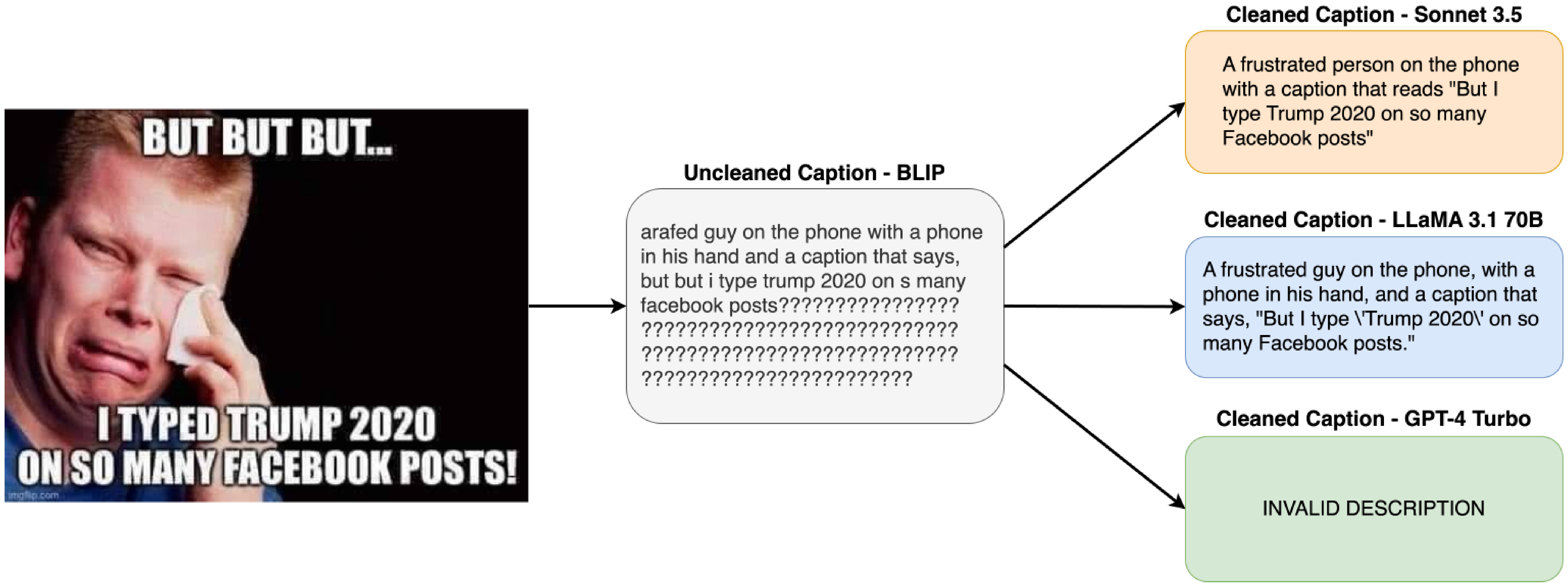
A more realistic output...

Uncleaned Caption - BLIP

arafed guy on the phone with a phone
in his hand and a caption that says,
but but i type trump 2020 on s many
facebook posts????????????????
????????????????????????????
????????????????????????????
????????????????????????



LLMs output:



But ...

why are they different?

Which method is better?

Coverage statistics:

Table 1: Caption Coverage Statistics for BLIP and GIT Models Using Sonnet 3.5, LLaMA 3.1 70B, and GPT-4 Turbo

Model	Set	Non-empty Captions (#)			Valid Captions (%)		
		Sonnet 3.5	LLaMA 3.1 70B	GPT-4 Turbo	Sonnet 3.5	LLaMA 3.1 70B	GPT-4 Turbo
BLIP	Train	6293	6993	4844	89.9%	99.9%	69.2%
	Dev	898	998	726	89.8%	99.8%	72.6%
	Test	895	1000	718	89.5%	100.0%	71.8%
GIT	Train	5075	6755	4872	72.5%	96.5%	69.6%
	Dev	676	958	698	67.6%	95.8%	69.8%
	Test	697	976	700	69.7%	97.6%	70.0%

GPT: conservative

LLaMA: loose

Sonnet: moderate

Experimental Setup:

Data:

Meme text, Meme Caption, Meme Caption Cleaned

Downstream model:

Google T5 (seq2seq, suits hierarchical labels).   ADD EXAMPLE meme

Baseline:

meme text only

Comparisons:

uncleaned vs llm-cleaned captions

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Results:

LLM	Set	Precision	Recall	F1
Baseline (No Caption)	Dev	73.95	56.72	64.20
	Test	67.84	47.35	55.77
Uncleaned Caption	Dev	75.83	56.52	64.77
	Test	68.47	49.03	57.14
Sonnet 3.5	Dev	74.83	58.69	65.78
	Test	65.65	50.75	57.25
LLaMA 70B	Dev	73.35	57.94	64.74
	Test	67.66	52.11	**58.87**
GPT-4 Turbo	Dev	74.76	58.83	**65.85**
	Test	69.82	49.55	57.96

Blip

LLM	Set	Precision	Recall	F1
Baseline (No Caption)	Dev	73.95	56.72	64.20
	Test	67.84	47.35	55.77
Uncleaned Caption	Dev	73.35	57.62	64.54
	Test	67.66	46.43	55.07
Sonnet 3.5	Dev	74.04	58.40	65.30
	Test	65.35	52.83	58.43
LLaMA 70B	Dev	73.33	58.98	65.37
	Test	67.71	51.79	**58.69**
GPT-4 Turbo	Dev	71.60	61.29	**66.04**
	Test	68.71	50.79	58.41

GIT

Insights

- Only one comparison showed a stat significant improvement
- GPT-4 is stricter (discards more, but cleaning more effectively)
- LLaMA retains most captions but may be permissive
- LLMs can modestly improve text quality for complex tasks
- Effect varies by LLM and source of noise

Is it Worth it?