On Zero-shot counterspeech generation by LLMs

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, Animesh Mukherjee

LREC-COLING 2024

⚠️ *This presentation contains material that is* ***offensive*** *or* ***hateful****; however this cannot be avoided owing to the nature of the work.*

Table of contents

1 Aim

2 Introduction

3 Experiments

4 Results

5 Conclusion

Table of contents

1 Aim

2 Introduction

3 Experiments

4 Results

5 Conclusion

AIM: Generate automated **counterspeech** once a **hate speech** is detected on a social media platform.

Table of contents

1 Aim

2 Introduction

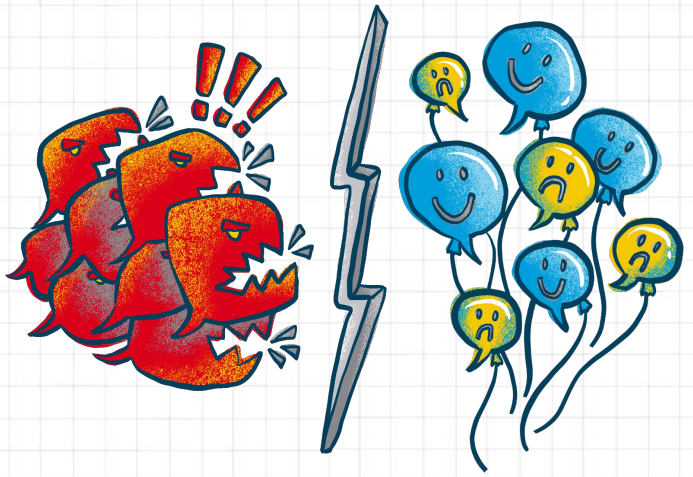
3 Experiments

4 Results

5 Conclusion

What is hate speech?

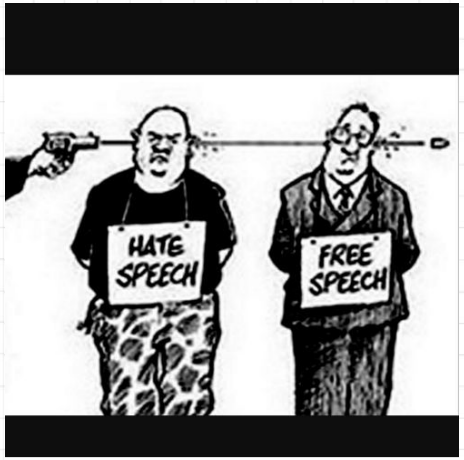
In common language, "hate speech" refers to offensive discourse targeting a group or an individual based on inherent characteristics (such as race, religion or gender) and **that may threaten societal peace and harmony.**



What is (usually) done after detecting hate speech?

- Deletion of posts
- Suspension of user accounts

What is done after detecting hate speech?



I may not always agree with what you say but I'll always support your right to say it

- Deletion of posts → Violates **freedom of speech**
- Suspension of user accounts → User may migrate to **other platforms** or create **new accounts**

Hence, these solutions are neither perfect nor permanent.

Counterspeech: a promising alternative

Counterspeech is an expression which aims to provide a counter argument to the hate speech with the aim of **de-escalating the conversation** and further influencing the **bystanders to act** and the **perpetrators** to **change their views**." (Benesch 2016)

Properties:

- It does not violate **freedom of speech.**
- It is flexible and responsive.

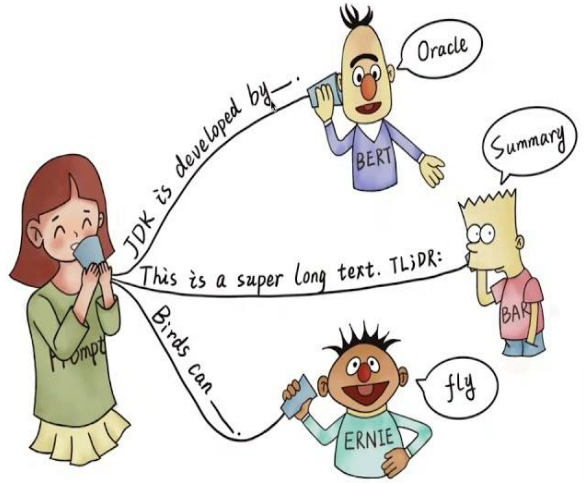
Curating counterspeech is difficult

- A **lot of manual effort** needed to write appropriate counterspeech given a hate speech.
- **Lack of expertise** of the counterspeech writers might lead to incorrect/incomplete counterspeech.
- Expert NGO workers are the best writers of counterspeech but they are **very expensive to hire**.
- Could be **often mentally taxing** for the writers due to the sheer volume of profane/abusive content in the hate speech posts
- Hate speech is **increasing at an unprecedented rate** → manually writing counterspeech for all of them is impossible

Alternative: **Use generative AI (e.g., LLMs) to generate counterspeech**

What is a prompt?

In the context of Natural Language Processing (NLP), a prompt is a piece of text used to guide a language model's generation of text. The language model is given a prompt as input, and it generates text that follows the prompt, often in the form of a continuation or a completion.



Basic framework for counterspeech generation

- Labeled data is scarce.
- How good are these LLMs in generating counterspeech in a zero-shot setting?
- Can we use prompt engineering to have LLMs generate counterspeech in a completely zero-shot setting?

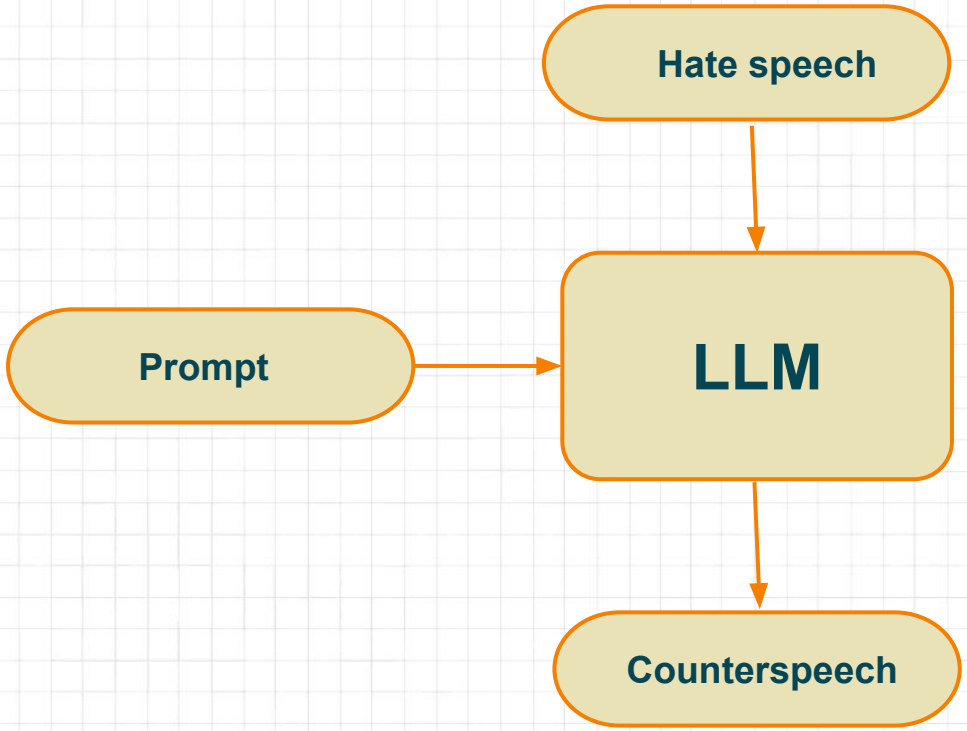


Table of contents

1 Aim

2 Introduction

3 Experiments

4 Results

5 Conclusion

Datasets

- **GAB:** Hate speech (6807), Counterspeech (40000)
- **CONAN-MT:** Hate speech (5000), Counterspeech (5000)
- **CONAN:** Hate speech (204), Counterspeech (3864)
- **Reddit:** Hate speech (2583), Counterspeech (7108)

Zero-shot prompting

- Completely zero-shot
  - Generated by providing only the hate speech as input to the model and asking it to generate a counterspeech

Type specific prompt variants

- Manual prompts
  - Prompts written manually using human knowledge assuming that they can be a good starting point for that counterspeech type
- Most frequent prompts
  - Generated by taking the most frequent beginning four words in the type-labeled dataset for each counterspeech type
- Cluster centered prompts
  - Generated prompts by clustering counterspeech of each type and then taking the center of the counterspeech (beginning four words) as prompts, assuming they represent the cluster

Manual prompts for each counterspeech type

- Prompts written manually using human knowledge thinking that they can be a good starting point for that counterspeech type

Counterspeech Type	Prompts
Affiliation	I also belong I know I have met
Denouncing	Saying this is not right Please do not say This is hate speech
Facts	This is a fact I know for a fact There is an evidence
Humor	This is funny You make me laugh The funny part is
Hypocrisy	In contradiction
Question	Do you really think that Are you aware of Where is the

Most frequent prompts for each counterspeech type

- Generated by taking the most frequent beginning four words in the type-labeled dataset for each counterspeech type

Counterspeech Type	Prompts
Affiliation	i am jewish and i am a muslim i really want to
Denouncing	this is not true how can you say have won at what
Facts	the vast majority of whilst a small number they are here because
Humor	i am really good I'd rather see them must be hard for
Hypocrisy	i do not think what about recent school why have we built
Question	have you thought about how can you say why do you think

Cluster centered prompts for each counterspeech type

- Generated prompts by clustering counterspeech of each type and then taking the center of the counterspeech (beginning four words) as prompts, assuming they represent the cluster

Counterspeech Type	Prompts
Affiliation	I feel very sorry I find it frightening This is nothing new
Denouncing	This is simply wrong This is not true How can you say
Facts	I suppose we have Anybody think that this In face at no
Humor	He should do standup How is this if I really cannot wait
Hypocrisy	You are not seriously This is why it People say they believe
Question	Can I ask you What do you mean Where is your evidence

Evaluation metrics

- **Generation metrics -** These metrics directly measure the similarity between the ground truth and the generated response for e.g. **gleu**, **bleurt** etc
- **Engagement prediction metrics -** These metrics measure the human feedback of the response given the hate speech for e.g. **width**, **height**, **upvotes**.
- **Quality metrics -** These metrics use different automatic classifiers to evaluate the **quality** of the response generated for e.g. **argument**, **counter argument**, **counterspeech** and **toxicity**

Table of contents

1 Aim

2 Introduction

3 Experiments

4 Results

5 Conclusion

Does counterspeech generation depend on model size in zero-shot setting?

Zero-shot Results (CONAN-MT)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.07	0.08	0.84	0.84	-1.13	0.15	0.54	0.26	0.24	65.21
DialoGPT-medium	0.07	0.08	0.84	0.84	-1.16	0.16	0.54	0.22	0.19	72.46
DialoGPT-large	0.07	0.09	0.83	0.83	-1.14	0.15	0.59	0.23	0.28	66.07
GPT2-small	0.06	0.11	0.82	0.84	-1.02	0.56	0.50	0.38	0.13	43.24
GPT2-medium	0.06	0.10	0.83	0.85	-1.05	0.51	0.49	0.36	0.18	44.25
GPT2-large	0.06	0.11	0.83	0.84	-1.04	0.48	0.47	0.36	0.19	43.27
Flan-T5-small	0.08	0.13	0.81	0.82	-0.94	0.40	0.56	0.48	0.18	61.21
Flan-T5-base	0.08	0.13	0.81	0.81	-0.90	0.43	0.49	0.47	0.21	60.65
Flan-T5-large	0.08	0.12	0.82	0.82	-0.96	0.41	0.46	0.43	0.18	58.08
ChatGPT	0.09	0.17	0.80	0.80	-0.53	0.95	0.64	0.51	0.15	29.89

Drop in quality - counterspeech (13%)

Zero-shot Results (CONAN)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.09	0.11	0.88	0.87	-1.21	0.15	0.58	0.20	0.31	60.67
DialoGPT-medium	0.09	0.11	0.88	0.87	-1.23	0.09	0.54	0.15	0.24	70.95
DialoGPT-large	0.09	0.12	0.86	0.86	-1.20	0.08	0.65	0.21	0.37	63.57
GPT2-small	0.08	0.15	0.85	0.86	-1.06	0.48	0.55	0.37	0.20	41.09
GPT2-medium	0.08	0.15	0.85	0.86	-1.06	0.34	0.54	0.38	0.23	44.65
GPT2-large	0.08	0.15	0.85	0.86	-1.08	0.43	0.51	0.35	0.21	43.73
Flan-T5-small	0.10	0.17	0.84	0.84	-1.86	0.33	0.56	0.40	0.26	61.59
Flan-T5-base	0.10	0.17	0.84	0.84	-1.84	0.33	0.52	0.44	0.25	53.48
Flan-T5-large	0.10	0.17	0.84	0.84	-0.98	0.31	0.58	0.42	0.22	55.32
ChatGPT	0.12	0.23	0.69	0.81	-0.63	0.89	0.64	0.44	0.23	32.05

Drop in quality - counterspeech (42%)

Zero-shot Results (CONAN-MT)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.07	0.08	0.84	0.84	-1.13	0.15	0.54	0.26	0.24	65.21
DialoGPT-medium	0.07	0.08	0.84	0.84	-1.16	0.16	0.54	0.22	0.19	72.46
DialoGPT-large	0.07	0.09	0.83	0.83	-1.14	0.15	0.59	0.23	0.28	66.07
GPT2-small	0.06	0.11	0.82	0.84	-1.02	0.56	0.50	0.38	0.13	43.24
GPT2-medium	0.06	0.10	0.83	0.85	-1.05	0.51	0.49	0.36	0.18	44.25
GPT2-large	0.06	0.11	0.83	0.84	-1.04	0.48	0.47	0.36	0.19	43.27
Flan-T5-small	0.08	0.13	0.81	0.82	-0.94	0.40	0.56	0.48	0.18	61.21
Flan-T5-base	0.08	0.13	0.81	0.81	-0.90	0.43	0.49	0.47	0.21	60.65
Flan-T5-large	0.08	0.12	0.82	0.82	-0.96	0.41	0.46	0.43	0.18	58.08
ChatGPT	0.09	0.17	0.80	0.80	-0.53	0.95	0.64	0.51	0.15	29.89

Drop in quality - counter argument (6-9%)

Zero-shot Results (CONAN-MT)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.07	0.08	0.84	0.84	-1.13	0.15	0.54	0.26	0.24	65.21
DialoGPT-medium	0.07	0.08	0.84	0.84	-1.16	0.16	0.54	0.22	0.19	72.46
DialoGPT-large	0.07	0.09	0.83	0.83	-1.14	0.15	0.59	0.23	0.28	66.07
GPT2-small	0.06	0.11	0.82	0.84	-1.02	0.56	0.50	0.38	0.13	43.24
GPT2-medium	0.06	0.10	0.83	0.85	-1.05	0.51	0.49	0.36	0.18	44.25
GPT2-large	0.06	0.11	0.83	0.84	-1.04	0.48	0.47	0.36	0.19	43.27
Flan-T5-small	0.08	0.13	0.81	0.82	-0.94	0.40	0.56	0.48	0.18	61.21
Flan-T5-base	0.08	0.13	0.81	0.81	-0.90	0.43	0.49	0.47	0.21	60.65
Flan-T5-large	0.08	0.12	0.82	0.82	-0.96	0.41	0.46	0.43	0.18	58.08
ChatGPT	0.09	0.17	0.80	0.80	-0.53	0.95	0.64	0.51	0.15	29.89

Increase in toxicity by 44 %

Zero-shot Results (Gab)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.05	0.07	0.87	0.86	-1.26	0.06	0.53	0.06	0.09	58.18
DialoGPT-medium	0.05	0.07	0.86	0.86	-1.26	0.08	0.55	0.06	0.09	57.00
DialoGPT-large	0.05	0.09	0.85	0.84	-1.28	0.07	0.56	0.06	0.09	59.25
GPT2-small	0.05	0.12	0.83	0.85	-1.37	0.31	0.53	0.19	0.15	58.16
GPT2-medium	0.05	0.12	0.84	0.85	-1.37	0.28	0.54	0.19	0.19	58.47
GPT2-large	0.05	0.12	0.83	0.85	-1.36	0.28	0.53	0.19	0.19	55.44
Flan-T5-small	0.06	0.11	0.84	0.84	-1.37	0.24	0.56	0.22	0.16	67.10
Flan-T5-base	0.06	0.11	0.84	0.83	-1.35	0.23	0.50	0.21	0.20	68.33
Flan-T5-large	0.06	0.11	0.84	0.83	-1.34	0.26	0.52	0.19	0.16	63.79
ChatGPT	0.08	0.17	0.64	0.80	-0.71	0.90	0.46	0.26	0.12	29.77

Increase in toxicity by 25%

Zero-shot Results (Reddit)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.05	0.06	0.87	0.88	-1.22	0.07	0.59	0.07	0.07	30.52
DialoGPT-medium	0.05	0.07	0.87	0.87	-1.21	0.08	0.55	0.06	0.07	58.24
DialoGPT-large	0.06	0.08	0.86	0.86	-1.25	0.08	0.61	0.07	0.07	62.26
GPT2-small	0.05	0.12	0.82	0.86	-1.34	0.36	0.57	0.21	0.12	55.06
GPT2-medium	0.05	0.12	0.83	0.86	-1.35	0.35	0.56	0.22	0.14	52.91
GPT2-large	0.05	0.12	0.83	0.86	-1.34	0.35	0.55	0.21	0.16	52.88
Flan-T5-small	0.06	0.12	0.83	0.84	-1.35	0.31	0.57	0.26	0.12	73.82
Flan-T5-base	0.06	0.11	0.84	0.84	-1.34	0.29	0.51	0.22	0.16	70.51
Flan-T5-large	0.06	0.11	0.84	0.84	-1.32	0.34	0.53	0.20	0.11	70.99
ChatGPT	0.08	0.17	67	81	-0.77	0.85	0.50	0.26	0.13	29.12

Increase in toxicity by 30 %

Zero-shot Results (Reddit)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.05	0.06	0.87	0.88	-1.22	0.07	0.59	0.07	0.07	30.52
DialoGPT-medium	0.05	0.07	0.87	0.87	-1.21	0.08	0.55	0.06	0.07	58.24
DialoGPT-large	0.06	0.08	0.86	0.86	-1.25	0.08	0.61	0.07	0.07	62.26
GPT2-small	0.05	0.12	0.82	0.86	-1.34	0.36	0.57	0.21	0.12	55.06
GPT2-medium	0.05	0.12	0.83	0.86	-1.35	0.35	0.56	0.22	0.14	52.91
GPT2-large	0.05	0.12	0.83	0.86	-1.34	0.35	0.55	0.21	0.16	52.88
Flan-T5-small	0.06	0.12	0.83	0.84	-1.35	0.31	0.57	0.26	0.12	73.82
Flan-T5-base	0.06	0.11	0.84	0.84	-1.34	0.29	0.51	0.22	0.16	70.51
Flan-T5-large	0.06	0.11	0.84	0.84	-1.32	0.34	0.53	0.20	0.11	70.99
ChatGPT	0.08	0.17	67	81	-0.77	0.85	0.50	0.26	0.13	29.12

Increase in readability by 100 %

Does counterspeech generation depend on model type in zero-shot setting?

Zero-shot Results (CONAN-MT)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.07	0.08	0.84	0.84	-1.13	0.15	0.54	0.26	0.24	65.21
DialoGPT-medium	0.07	0.08	0.84	0.84	-1.16	0.16	0.54	0.22	0.19	72.46
DialoGPT-large	0.07	0.09	0.83	0.83	-1.14	0.15	0.59	0.23	0.28	66.07
GPT2-small	0.06	0.11	0.82	0.84	-1.02	0.56	0.50	0.38	0.13	43.24
GPT2-medium	0.06	0.10	0.83	0.85	-1.05	0.51	0.49	0.36	0.18	44.25
GPT2-large	0.06	0.11	0.83	0.84	-1.04	0.48	0.47	0.36	0.19	43.27
Flan-T5-small	0.08	0.13	0.81	0.82	-0.94	0.40	0.56	0.48	0.18	61.21
Flan-T5-base	0.08	0.13	0.81	0.81	-0.90	0.43	0.49	0.47	0.21	60.65
Flan-T5-large	0.08	0.12	0.82	0.82	-0.96	0.41	0.46	0.43	0.18	58.08
ChatGPT	0.09	0.17	0.80	0.80	-0.53	0.95	0.64	0.51	0.15	29.89

Flan-T5 has the best gleu, meteor and bleurt for synthetic datasets

Zero-shot Results (CONAN-MT)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.07	0.08	0.84	0.84	-1.13	0.15	0.54	0.26	0.24	65.21
DialoGPT-medium	0.07	0.08	0.84	0.84	-1.16	0.16	0.54	0.22	0.19	72.46
DialoGPT-large	0.07	0.09	0.83	0.83	-1.14	0.15	0.59	0.23	0.28	66.07
GPT2-small	0.06	0.11	0.82	0.84	-1.02	0.56	0.50	0.38	0.13	43.24
GPT2-medium	0.06	0.10	0.83	0.85	-1.05	0.51	0.49	0.36	0.18	44.25
GPT2-large	0.06	0.11	0.83	0.84	-1.04	0.48	0.47	0.36	0.19	43.27
Flan-T5-small	0.08	0.13	0.81	0.82	-0.94	0.40	0.56	0.48	0.18	61.21
Flan-T5-base	0.08	0.13	0.81	0.81	-0.90	0.43	0.49	0.47	0.21	60.65
Flan-T5-large	0.08	0.12	0.82	0.82	-0.96	0.41	0.46	0.43	0.18	58.08
ChatGPT	0.09	0.17	0.80	0.80	-0.53	0.95	0.64	0.51	0.15	29.89

GPT-2 has the best counterspeech score.

Zero-shot Results (Gab)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.05	0.07	0.87	0.86	-1.26	0.06	0.53	0.06	0.09	58.18
DialoGPT-medium	0.05	0.07	0.86	0.86	-1.26	0.08	0.55	0.06	0.09	57.00
DialoGPT-large	0.05	0.09	0.85	0.84	-1.28	0.07	0.56	0.06	0.09	59.25
GPT2-small	0.05	0.12	0.83	0.85	-1.37	0.31	0.53	0.19	0.15	58.16
GPT2-medium	0.05	0.12	0.84	0.85	-1.37	0.28	0.54	0.19	0.19	58.47
GPT2-large	0.05	0.12	0.83	0.85	-1.36	0.28	0.53	0.19	0.19	55.44
Flan-T5-small	0.06	0.11	0.84	0.84	-1.37	0.24	0.56	0.22	0.16	67.10
Flan-T5-base	0.06	0.11	0.84	0.83	-1.35	0.23	0.50	0.21	0.20	68.33
Flan-T5-large	0.06	0.11	0.84	0.83	-1.34	0.26	0.52	0.19	0.16	63.79
ChatGPT	0.08	0.17	0.64	0.80	-0.71	0.90	0.46	0.26	0.12	29.77

Different models are best in terms of different generational metrics

Zero-shot Results (Gab)

Model	gleu	met	div	nov	blrt	cs	c_arg	arg	tox	fre
DialoGPT-small	0.05	0.07	0.87	0.86	-1.26	0.06	0.53	0.06	0.09	58.18
DialoGPT-medium	0.05	0.07	0.86	0.86	-1.26	0.08	0.55	0.06	0.09	57.00
DialoGPT-large	0.05	0.09	0.85	0.84	-1.28	0.07	0.56	0.06	0.09	59.25
GPT2-small	0.05	0.12	0.83	0.85	-1.37	0.31	0.53	0.19	0.15	58.16
GPT2-medium	0.05	0.12	0.84	0.85	-1.37	0.28	0.54	0.19	0.19	58.47
GPT2-large	0.05	0.12	0.83	0.85	-1.36	0.28	0.53	0.19	0.19	55.44
Flan-T5-small	0.06	0.11	0.84	0.84	-1.37	0.24	0.56	0.22	0.16	67.10
Flan-T5-base	0.06	0.11	0.84	0.83	-1.35	0.23	0.50	0.21	0.20	68.33
Flan-T5-large	0.06	0.11	0.84	0.83	-1.34	0.26	0.52	0.19	0.16	63.79
ChatGPT	0.08	0.17	0.64	0.80	-0.71	0.90	0.46	0.26	0.12	29.77

While FlanT5 is great in terms of counter speech score, it gets worse in terms of toxicity

Summary

- **Does counterspeech generation depend on model size in zero-shot setting?**
  Ans: No, not necessarily. Counter-intuitively we notice the performance drop in some cases.

- **Does counterspeech generation depend on model type in zero-shot setting?**
  Ans: Yes, the inherent properties of the model are important and also dependant on the dataset. We can select a model based on our use-case.

Counterspeech types: genesis of the other prompt variants

- Counterspeech type: Roughly the **strategies** of the counterspeech.
  E.g. affiliation, denouncing, facts, humour, hypocrisy, questions etc...

- A **fact** type counterspeech could be – "a quick google search shows that your stats are plain wrong about everything and most Global Terrorism is motivated by politics and not religion. That is why no one is talking about it."

- An **affiliation** type counterspeech could be - "I would never laugh at your murder. I would never taunt your grieving. I would never mock your fight for equality."

- A **denouncing** type counterspeech could be- "@MissAmerica sorry for being rude and 'racist' and calling you a Arab please tweet back so everyone will know its real."

Counterspeech types: genesis of the other prompt variants

- A **humor** type counterspeech could be “They use this guy for 95% of racial discrimination experiments. he must be really good at it lmao.”
- A **hypocrisy** type counterspeech is “If bribery in Islam is not allowed then why was Naik convicted of paying people to convert to Islam. Even Mohamed used to bribe enemies he wanted to convert..”
- An example for **question** type counterspeech is “What is wrong with accepting and understanding people of different backgrounds?”

Counterspeech classification

- Trained this classifier to assess whether our models are creating correct type of counterspeech given the prompt type
- Utilised the WikiLingua Dataset (English texts) and the dataset from “Thou Shalt Not Hate” to fine tune our BERT model
- Allocated only 70% of this data for fine-tuning the classification model
- Reserved the remaining 30% for generating most frequent and cluster-centered prompts
- Done to avoid any type of bias resulting from data leakage
- The classifier model achieved a classification accuracy of approximately 76% and a macro F1 score of 0.80.

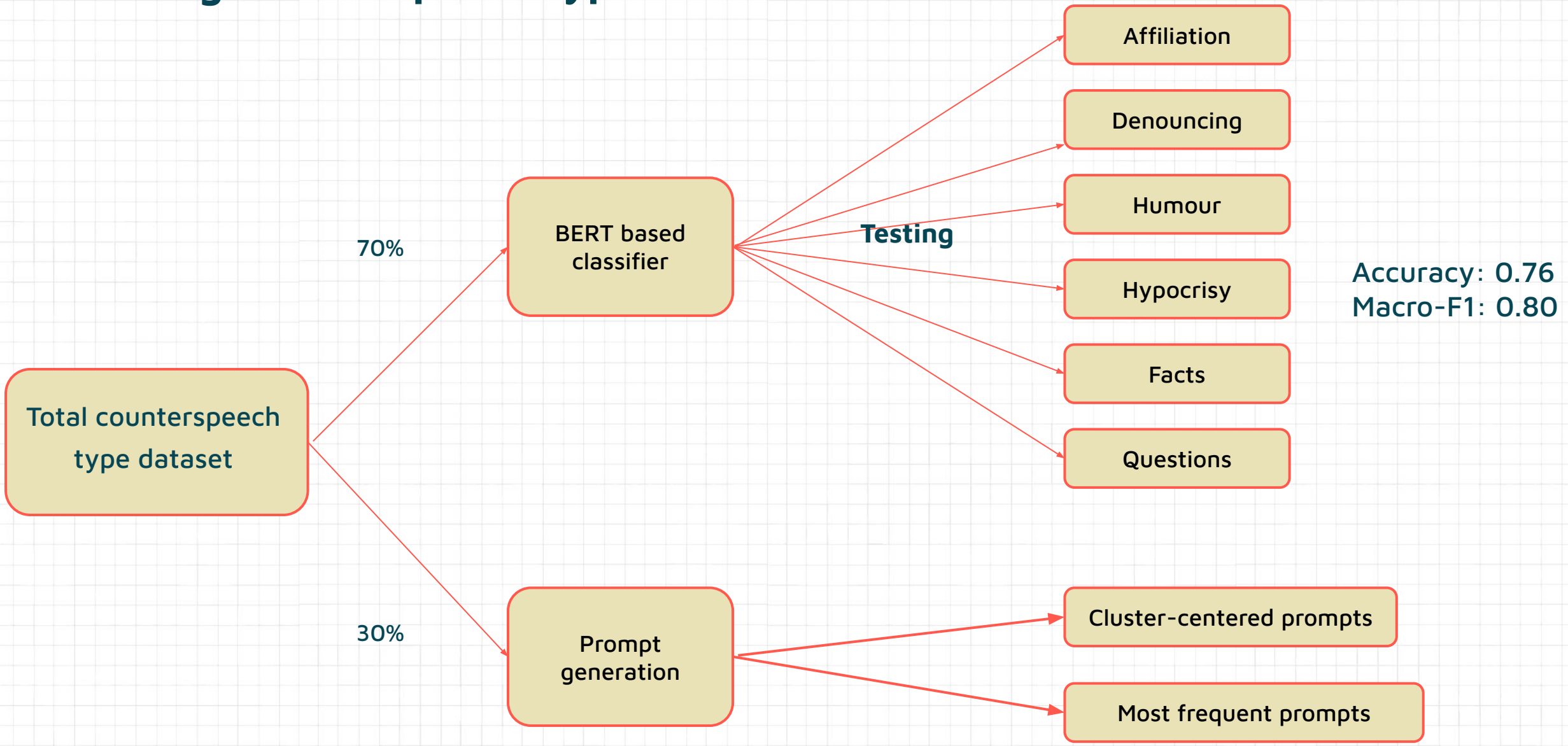
Counterspeech type Datasets:

Also used the following type annotated datasets

- Dataset used in “Multilingual Counter Narrative Type Classification” and “Thou Shalt Not Hate: Countering Online Hate Speech”

- These datasets were used for training the classification model as well as in generation of prompts

Obtaining counterspeech types



Accuracy: 0.76
Macro-F1: 0.80

Prompt-variant vs type results

CONAN_MT							
Model	Prompt	Affiliation	Denouncing	Facts	Humor	Hypocrisy	Questions
GPT-2	base	0.04	0.04	0.60	0.03	0.29	0.00
	manual	0.06	0.40	0.78	0.20	0.21	0.01
	freq	0.06	0.07	0.77	0.10	0.21	0.05
	cluster	0.46	0.34	0.72	0.07	0.43	0.03
DialoGPT	base	0.04	0.10	0.29	0.19	0.39	0.00
	manual	0.10	0.45	0.46	0.46	0.51	0.01
	freq	0.10	0.15	0.44	0.34	0.50	0.08
	cluster	0.38	0.41	0.42	0.29	0.49	0.03
Flan-T5	base	0.04	0.06	0.73	0.03	0.14	0.00
	manual	0.15	0.23	0.81	0.15	0.10	0.00
	freq	0.26	0.06	0.80	0.07	0.21	0.00
	cluster	0.20	0.18	0.74	0.10	0.23	0.00
ChatGPT	base	0.02	0.22	0.75	0.00	0.01	0.00
	manual	0.03	0.40	0.81	0.00	0.02	0.00
	freq	0.51	0.31	0.94	0.00	0.06	0.01
	cluster	0.27	0.46	0.77	0.00	0.09	0.01

~0

"Questions" type is very hard to generate , ChatGPT fails at generating "Humor" type

Prompt-variant vs type results

CONAN							
Model	Prompt	Affiliation	Denouncing	Facts	Humor	Hypocrisy	Questions
GPT-2	base	0.02	0.04	0.65	0.01	0.27	0.00
	manual	0.05	0.37	0.81	0.13	0.19	0.00
	freq	0.04	0.07	0.79	0.05	0.18	0.04
	cluster	0.44	0.33	0.76	0.04	0.39	0.02
DialoGPT	base	0.02	0.09	0.28	0.20	0.41	0.00
	manual	0.07	0.44	0.42	0.46	0.52	0.01
	freq	0.07	0.16	0.42	0.37	0.50	0.07
	cluster	0.36	0.41	0.35	0.32	0.50	0.03
Flan-T5	base	0.02	0.08	0.73	0.02	0.12	0.00
	manual	0.10	0.23	0.79	0.14	0.12	0.00
	freq	0.18	0.07	0.80	0.05	0.21	0.00
	cluster	0.16	0.17	0.73	0.08	0.23	0.00
ChatGPT	base	0.04	0.64	0.28	0.00	0.10	0.00
	manual	0.02	0.39	0.93	0.00	0.02	0.00
	freq	0.52	0.30	0.96	0.00	0.06	0.00
	cluster	0.26	0.43	0.87	0.01	0.10	0.01

ChatGPT outperforms other models for “Affiliation”, “Denouncing” and “Facts” type

Prompt-variant vs type results

REDDIT							
Model	Prompt	Affiliation	Denouncing	Facts	Humor	Hypocrisy	Questions
GPT-2	base	0.08	0.07	0.27	0.17	0.41	0.00
	manual	0.20	0.48	0.51	0.40	0.48	0.01
	freq	0.20	0.16	0.50	0.26	0.50	0.08
	cluster	0.47	0.41	0.47	0.21	0.50	0.03
DialoGPT	base	0.03	0.07	0.14	0.47	0.30	0.00
	manual	0.08	0.43	0.30	0.75	0.52	0.01
	freq	0.08	0.14	0.29	0.64	0.52	0.07
	cluster	0.33	0.38	0.29	0.52	0.41	0.02
Flan-T5	base	0.09	0.12	0.26	0.23	0.30	0.00
	manual	0.17	0.27	0.44	0.34	0.30	0.00
	freq	0.23	0.12	0.41	0.25	0.30	0.01
	cluster	0.19	0.21	0.43	0.24	0.31	0.01
ChatGPT	base	0.07	0.39	0.44	0.00	0.10	0.00
	manual	0.07	0.60	0.75	0.02	0.10	0.00
	freq	0.54	0.48	0.75	0.00	0.13	0.01
	cluster	0.35	0.57	0.57	0.01	0.12	0.01

DialoGPT is better for generating “Humor” and “Hypocrisy” type

Prompt-variant vs type results

GAB							
Model	Prompt	Affiliation	Denouncing	Facts	Humor	Hypocrisy	Questions
GPT-2	base	0.08	0.09	0.26	0.18	0.40	0.00
	manual	0.23	0.49	0.51	0.40	0.47	0.01
	freq	0.23	0.19	0.48	0.28	0.46	0.07
	cluster	0.50	0.44	0.45	0.20	0.49	0.03
DialoGPT	base	0.03	0.08	0.12	0.47	0.29	0.00
	manual	0.09	0.46	0.27	0.73	0.52	0.01
	freq	0.09	0.16	0.27	0.64	0.53	0.07
	cluster	0.33	0.41	0.26	0.52	0.39	0.02
Flan-T5	base	0.10	0.12	0.25	0.25	0.29	0.00
	manual	0.17	0.28	0.43	0.37	0.29	0.00
	freq	0.23	0.11	0.41	0.28	0.29	0.00
	cluster	0.19	0.23	0.43	0.26	0.33	0.00
ChatGPT	base	0.02	0.09	0.88	0.00	0.01	0.00
	manual	0.07	0.71	0.62	0.01	0.07	0.00
	freq	0.52	0.66	0.65	0.00	0.09	0.00
	cluster	0.30	0.69	0.52	0.00	0.07	0.00

Manual prompts perform best for “Denouncing”, “Facts” and “Humor” type

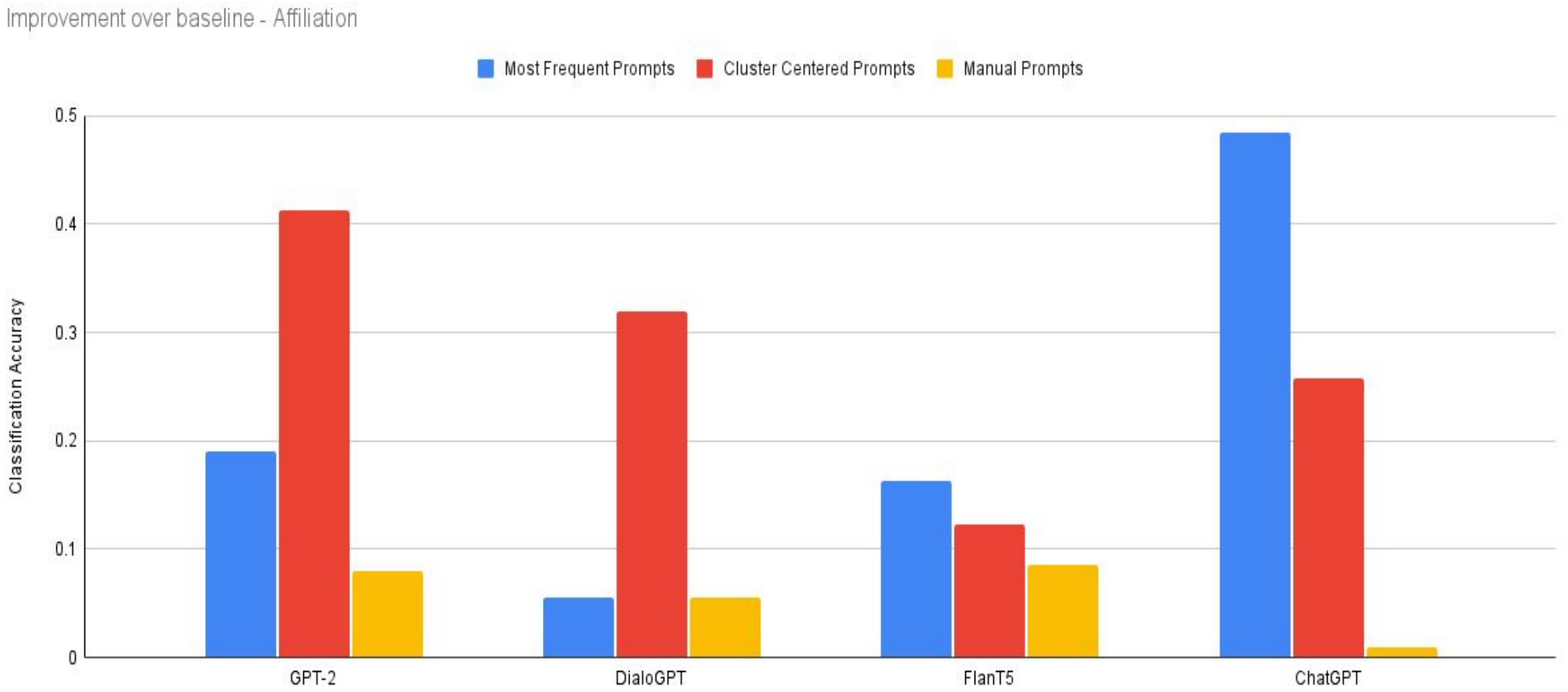
Summary

- A significant gain in scores over the baseline results → **prompt engineering is really helpful**
- Different models excel in generating different types of counterspeech. For example, ChatGPT performs better in generating counterspeech for Affiliation, Denouncing, and Facts types, while D-GPT excels in generating humor and hypocrisy type counterspeech
- For most of types, either **manual prompts (Denouncing, Facts, Humour)** or **Cluster centered prompts (Affiliation and Hypocrisy)** are most suitable.
- Surprisingly, ChatGPT was not able to generate "**Humour**" type at all
- "**Question**" was the hardest type of counterspeech to generate

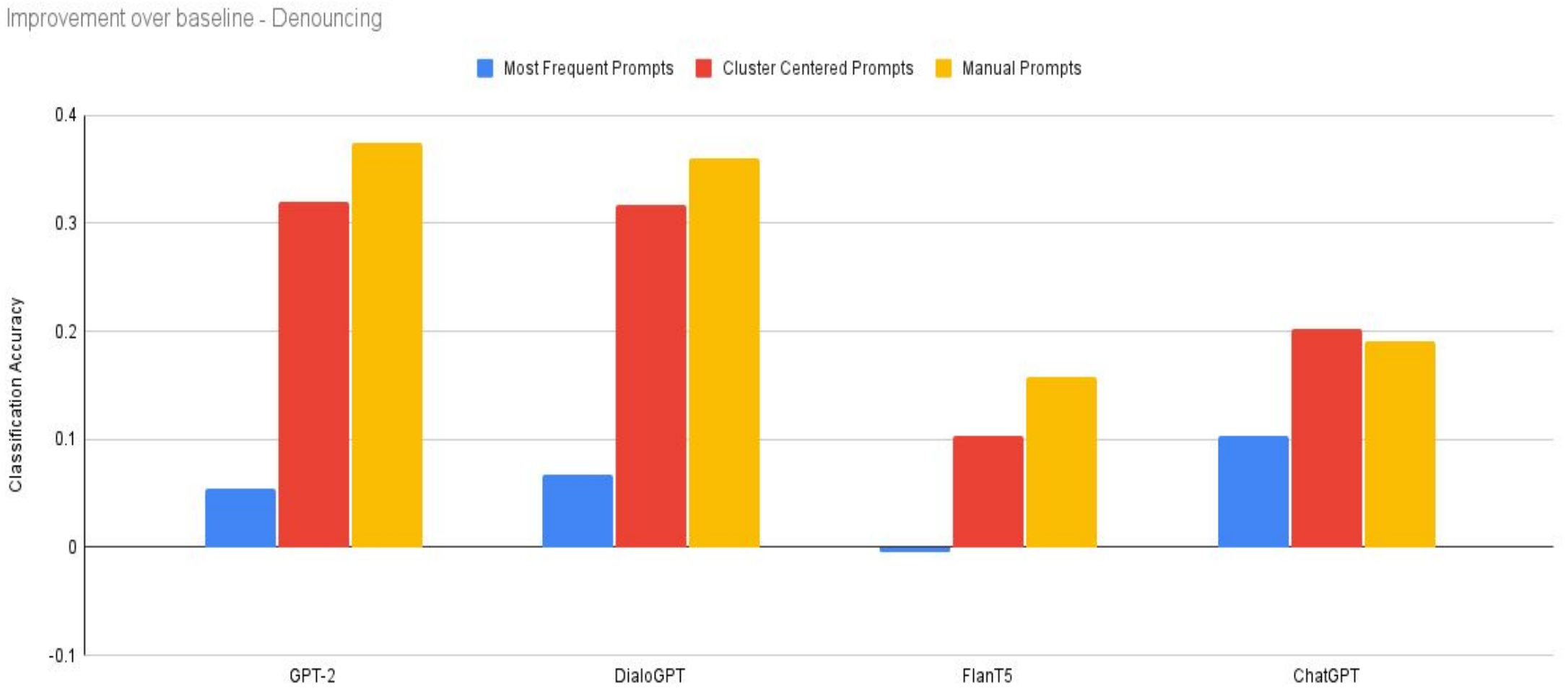
Comparative study

- Compared all three prompting methods with the baseline scores
- Calculated the difference between the score of each prompting strategy and the respective baseline score
- To obtain a single value per model across datasets, we averaged the values across the three variations (small, medium, and large) and all four datasets.

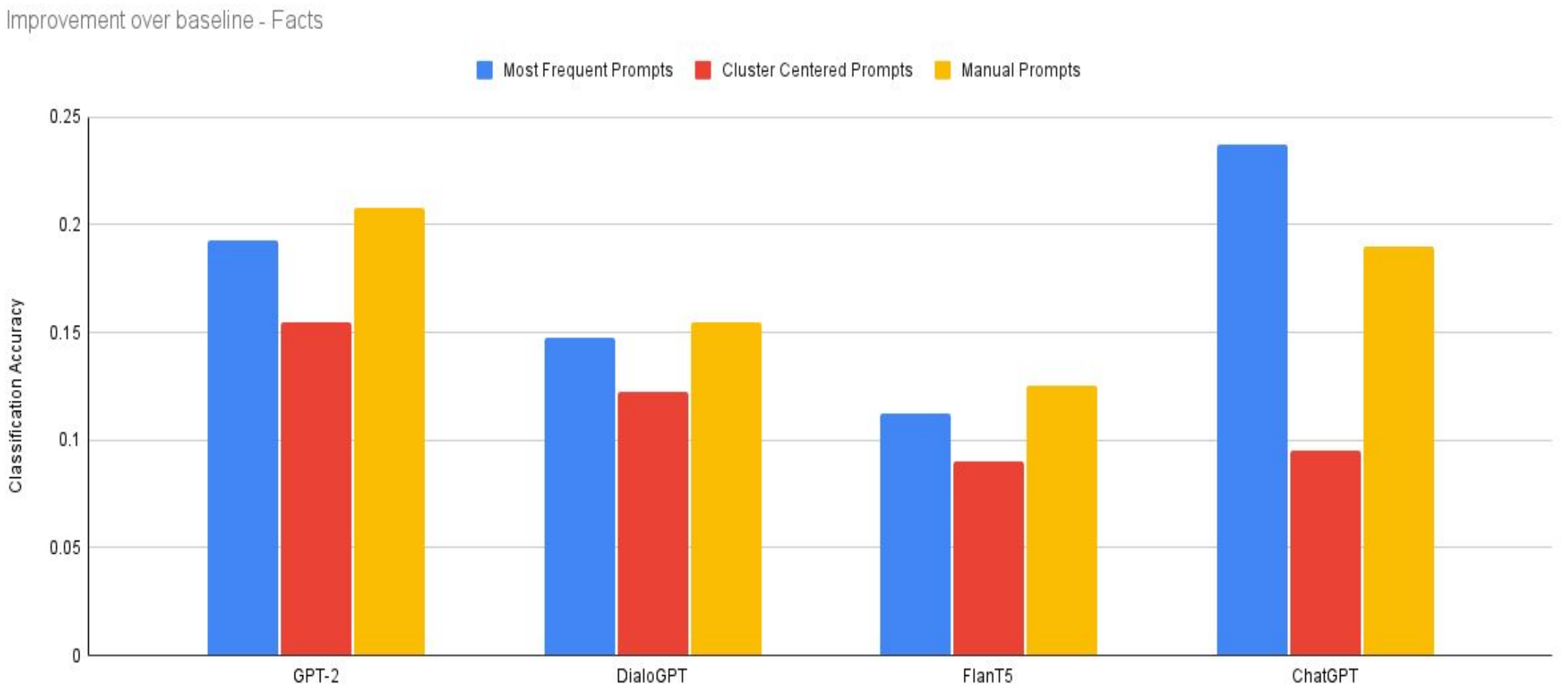
Affiliation



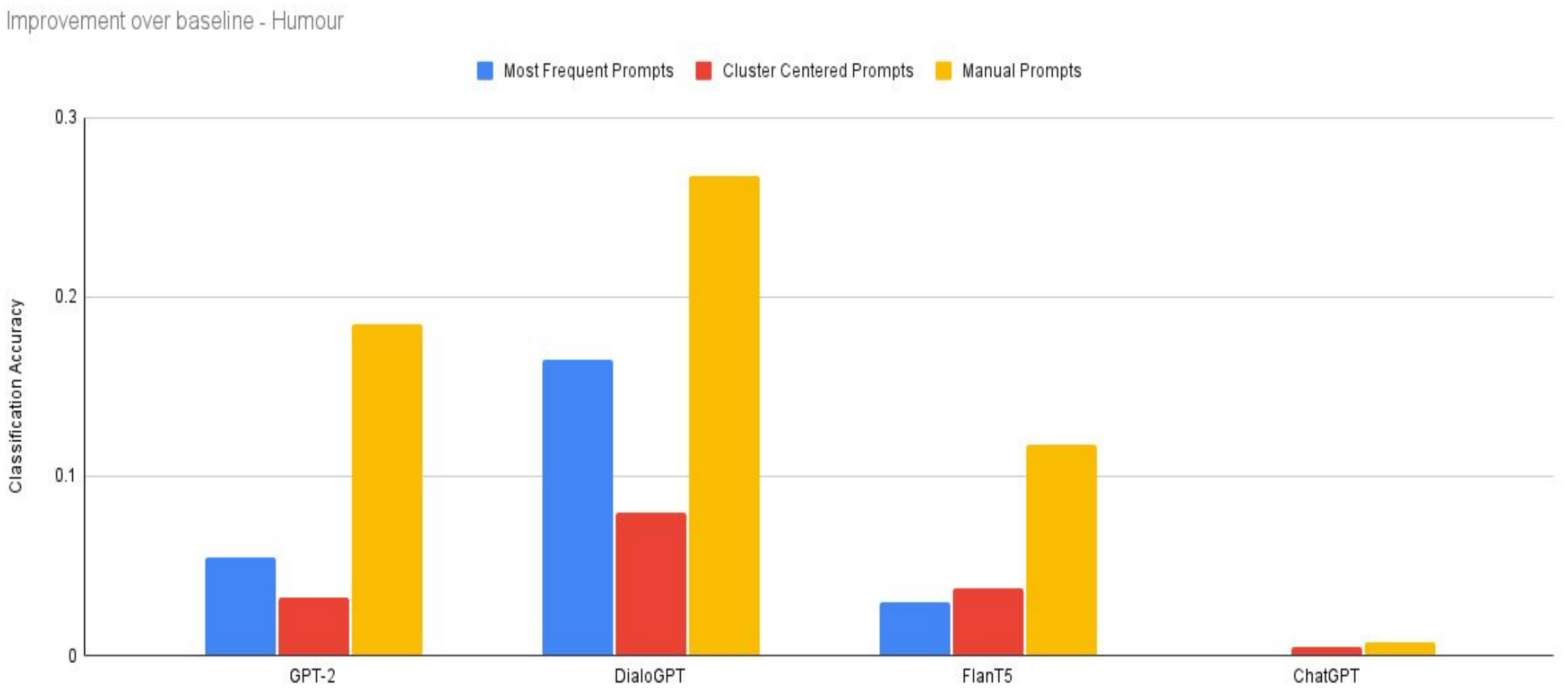
Denouncing


Improvement over baseline - Denouncing
Most Frequent Prompts
Cluster Centered Prompts
Manual Prompts
Classification Accuracy
0.4
0.3
0.2
0.1
0
-0.1
GPT-2
DialoGPT
FlanT5
ChatGPT

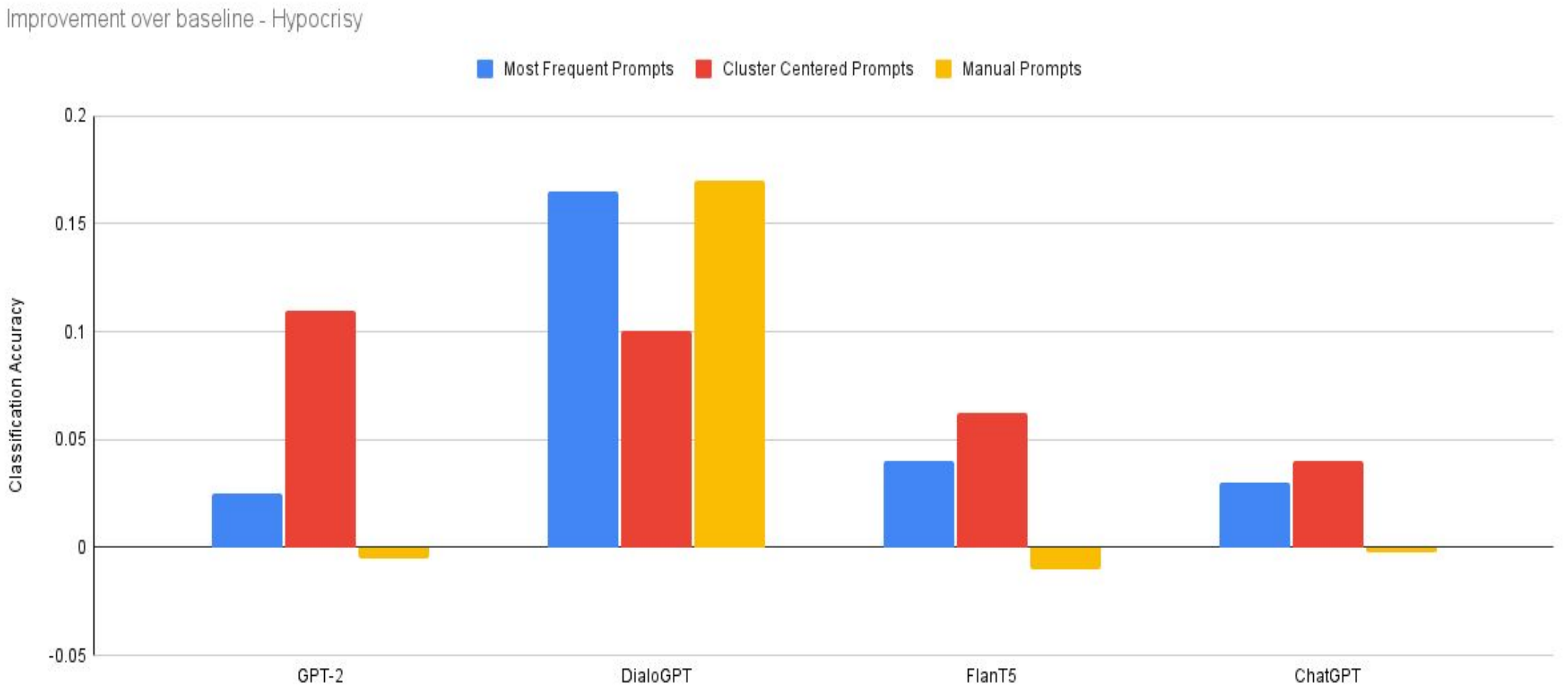
Facts


Improvement over baseline - Facts
Most Frequent Prompts
Cluster Centered Prompts
Manual Prompts
Classification Accuracy
0.25
0.2
0.15
0.1
0.05
0
GPT-2
DialoGPT
FlanT5
ChatGPT

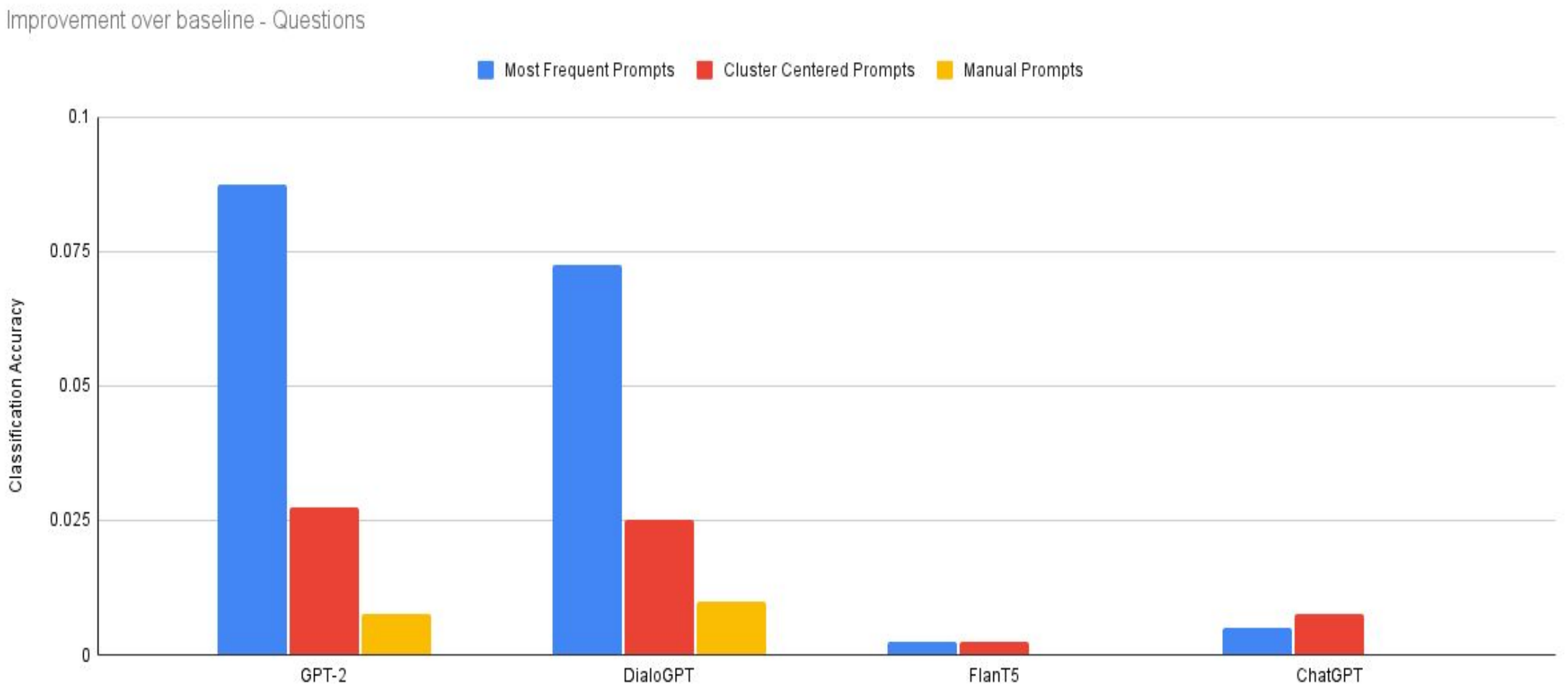
Humour


Improvement over baseline - Humour
Most Frequent Prompts
Cluster Centered Prompts
Manual Prompts
Classification Accuracy
0.3
0.2
0.1
0
GPT-2
DialoGPT
FlanT5
ChatGPT

Hypocrisy


Improvement over baseline - Hypocrisy
Most Frequent Prompts
Cluster Centered Prompts
Manual Prompts
Classification Accuracy
0.2
0.15
0.1
0.05
0
-0.05
GPT-2
DialoGPT
FlanT5
ChatGPT

Question


Improvement over baseline - Questions
Most Frequent Prompts
Cluster Centered Prompts
Manual Prompts
Classification Accuracy
0.1
0.075
0.05
0.025
0
GPT-2
DialoGPT
FlanT5
ChatGPT

Comparative study

- The figure clearly indicates that creating a pipeline might be necessary to generate effective counterspeech of a specific type, depending on the type itself

Counterspeech Type	Model	Prompt Type
Affiliation	ChatGPT	Most frequent prompts
Denouncing	GPT-2	Manual prompts
Facts	ChatGPT	Most frequent prompts
Humour	DialoGPT	Manual prompts
Hypocrisy	DialoGPT	Manual prompts
Questions	GPT-2	Most frequent prompts

Examples

Model	Hatespeech	Counterspeech
Flan-T5	Muslims have no place in civilised Europe. They should go home.	We shall not discriminate on the grounds of race, religion, colour, culture or national origin. Our society defends and protects its own dignity.
DGPT	We are being invaded by Muslims. They are taking over our countries	lol I hope we all get this reference.The irony is lost at the end of that sentence isn't it?
ChatGPT	If you are a Jew you are for sure involved in 9/11 planning. This is all about Race and 'Enlightenment'. You are all guilty until proven innocent.	It is unfair and untrue to suggest that all Jews were involved in the planning of 9/11...

Table of contents

1 Aim

2 Introduction

3 Experiments

4 Results

5 Conclusion

- The increase in size does not necessarily imply better counterspeech generation.
- Prompting helps in the better generation of a particular type of counterspeech.
- Different prompting methods can be used based on the type of counterspeech to be generated

Thanks!

Punyajoy Saha

Aalok Agrawal

Abhik Jana

Chris Biemann

Animesh Mukherjee

Send your questions to punyajoys@iitkgp.ac.in

Find more about us here !

GITHUB