Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Do Large Language Models Understand Mansplaining? *Well, actually...*

Carla Pérez Almendros and Jose Camacho-Collados

LREC-COLING 2024

Mansplaining: Definition

“The explanation of something by a man, typically to a woman, in a manner regarded as condescending or patronizing”

Oxford Languages

Research question

Do LLMs understand mansplaining?

- ➔ Zero-shot setting
- ➔ Implicitly and explicitly

Well, actually….

Data collection and analysis

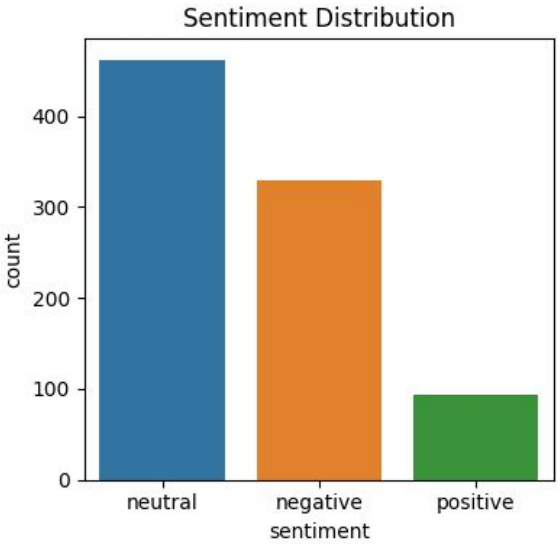
Well actually... The Corpus

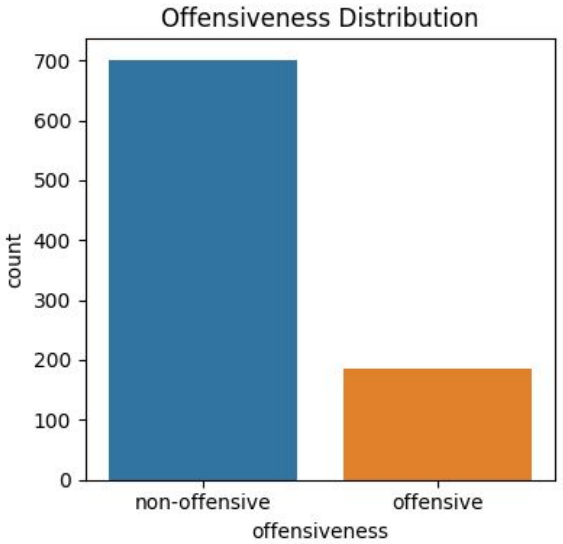
Collected from Reddit (thread on mansplaining stories)

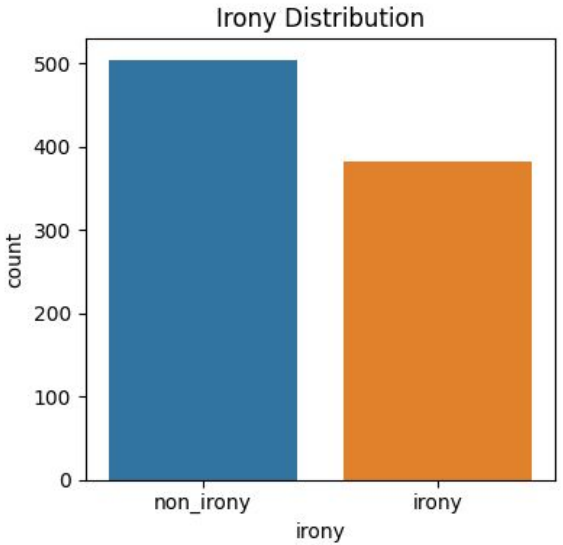
Manually curated (filtering stories, removing comments, etc.)

After curation: 886 mansplaining stories

The *Well, actually* corpus: Overview

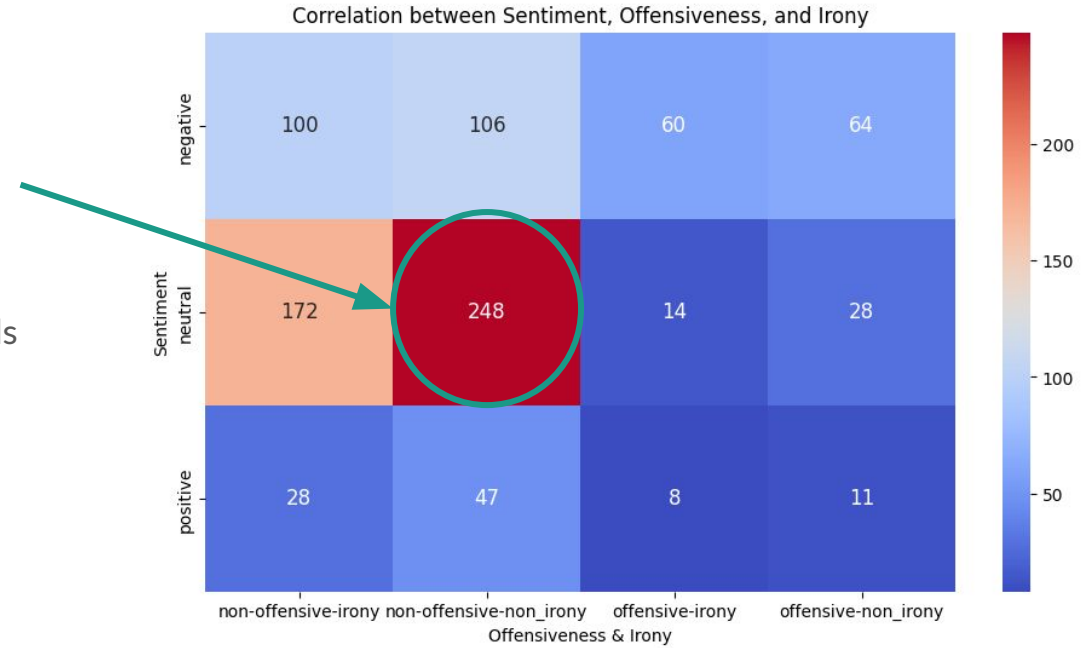






The *Well, actually* corpus: Overview

Most stories are classified as **non-offensive**, **non-ironic** and with **neutral** sentiment by automatic NLP models



The *Well, actually* corpus: Topic analysis

	Topic	Stories
1	Author experience and frustration	109
2	Gender roles and stereotypes	71
3	Mansplaining in tech	58
4	Work experience and Education	56
5	Food use and preparation	43
6	Gender bias at the work place	42
7	Arts, literature and movies	38
8	Pregnancy, childbirth & breastfeeding	33
9	Cars and motor	30
10	Sexual relationships	29

The *Well, actually* corpus: Topic analysis

	Topic	Stories
1	Author experience and frustration	109
2	Gender roles and stereotypes	71
3	Mansplaining in tech	58
4	Work experience and Education	56
5	Food use and preparation	43
6	Gender bias at the work place	42
7	Arts, literature and movies	38
8	Pregnancy, childbirth & breastfeeding	33
9	Cars and motor	30
10	Sexual relationships	29

The *Well, actually* corpus: Topic analysis

	Topic	Stories
1	Author experience and frustration	109
2	Gender roles and stereotypes	71
3	Mansplaining in tech	58
4	Work experience and Education	56
5	Food use and preparation	43
6	Gender bias at the work place	42
7	Arts, literature and movies	38
8	Pregnancy, childbirth & breastfeeding	33
9	Cars and motor	30
10	Sexual relationships	29

Evaluation

Experimental setting: LLMs

➔ ChatGPT (3.5-Turbo)

➔ LLaMA 2 (13B and 70B)

Zero-shot setting in both cases

Questions/prompts

Explicit questions:

- ➔ **Q1:** *Can you identify gender bias in this situation?*
- ➔ **Q2:** *Is this a case of mansplaining?*

Questions/prompts

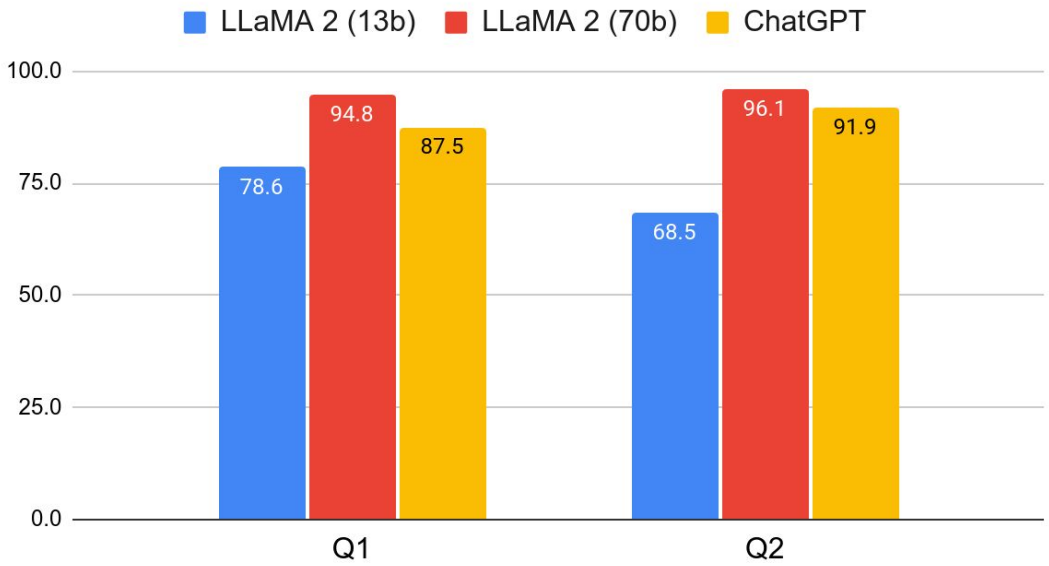
Explicit questions:

- ➔ **Q1:** *Can you identify gender bias in this situation?*
- ➔ **Q2:** *Is this a case of mansplaining?*

Implicit questions:

- ➔ **Q3:** *What can you infer from this situation?*
- ➔ **Q4:** *Can you identify up to five topics that appear in the text?*

Results (Explicit questions - accuracy)

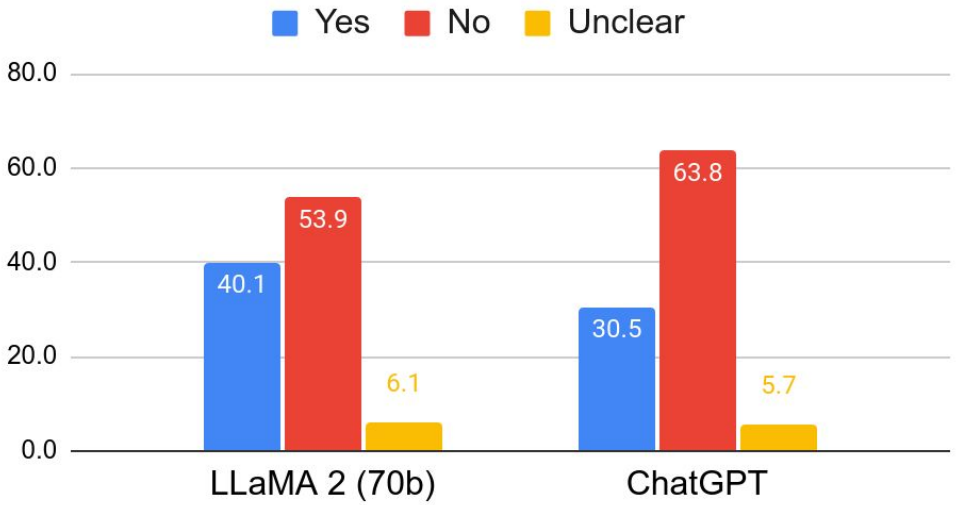


Can you identify gender bias in this situation?

Is this a case of mansplaining?

Results Q3 (implicit, % after manual check)

Q3: *What can you infer from this situation?*



Mansplaining or gender discrimination inferred?

Results Q4 (implicit, after manual check)

Q4: *Can you identify up to five topics that appear in the text?*

Percentage of stories without topics related to mansplaining or gender discrimination:

- ➔ **ChatGPT:** 55.7%
- ➔ **LLaMA 2 (70b):** 45.8%

➡️ In around half of all instances, models could not identify mansplaining or discrimination when asked to list five topics!

Conclusion

Identifying mansplaining situations is not an easy task for LLMs

To a certain extent, LLMs can identify these situations when asked directly, but **not if asked implicitly**!

Conclusion

Identifying mansplaining situations is not an easy task for LLMs

To a certain extent, LLMs can identify these situations when asked directly, but **not if asked implicitly**!

📄 **Release:** The curated *Well, actually…* corpus is openly available in Hugging Face 🤗

🌐 https://huggingface.co/datasets/carlaperez/Well_actually_mansplaining