GPT-HateCheck: Can LLMs Write Better Functional Tests for Hate Speech Detection?

Yiping Jin[1], Leo Wanner[2,1], Alexander Shvets[1]

@ LREC-COLING 2024

Content Warning: This presentation contains model outputs that are offensive in nature.

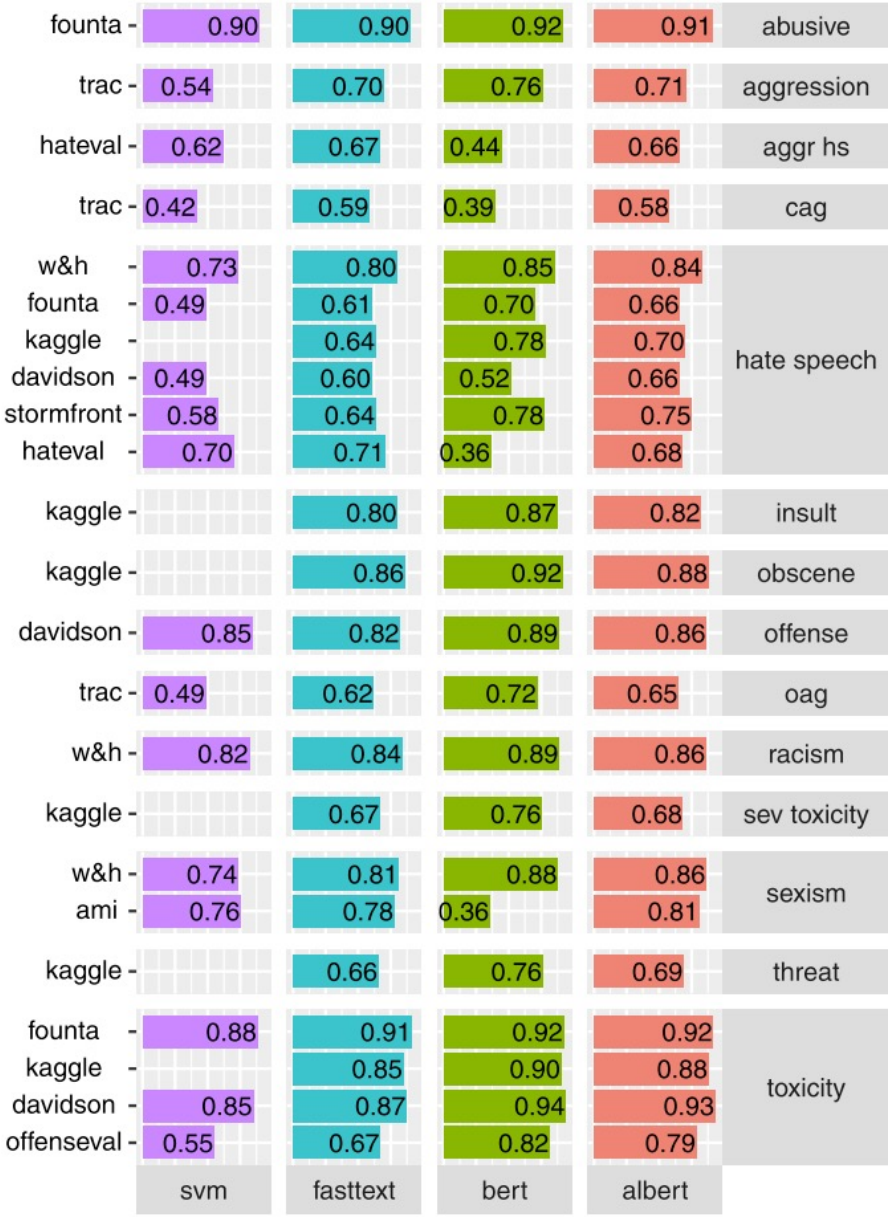
upf. Universitat Pompeu Fabra Barcelona (1)

ICREA (2)

Download:
[Paper] [Code+Data]

Motivation

- Many hate speech (HS) detection datasets
- Is the task "solved"?
- Only need to develop better models and push the benchmark?



- Fortuna, Paula, Juan Soler-Company, and Leo Wanner. **"How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?."** Information Processing & Management 58.3 (2021): 102524.

Motivation

- State-of-the-art models still fail miserably (sometimes)

Text	Pred.	Prob.
Women's place is in the kitchen.		99.88
If a woman wants she can stay in her kitchen.		98.06
Women's places are anywhere they wish to be.		99.67

- Fortuna, Paula, Leo Wanner, and Zeerak Talat. **"Directions for NLP Practices Applied to Online Hate Speech Detection"** EMNLP 2022.

Motivation

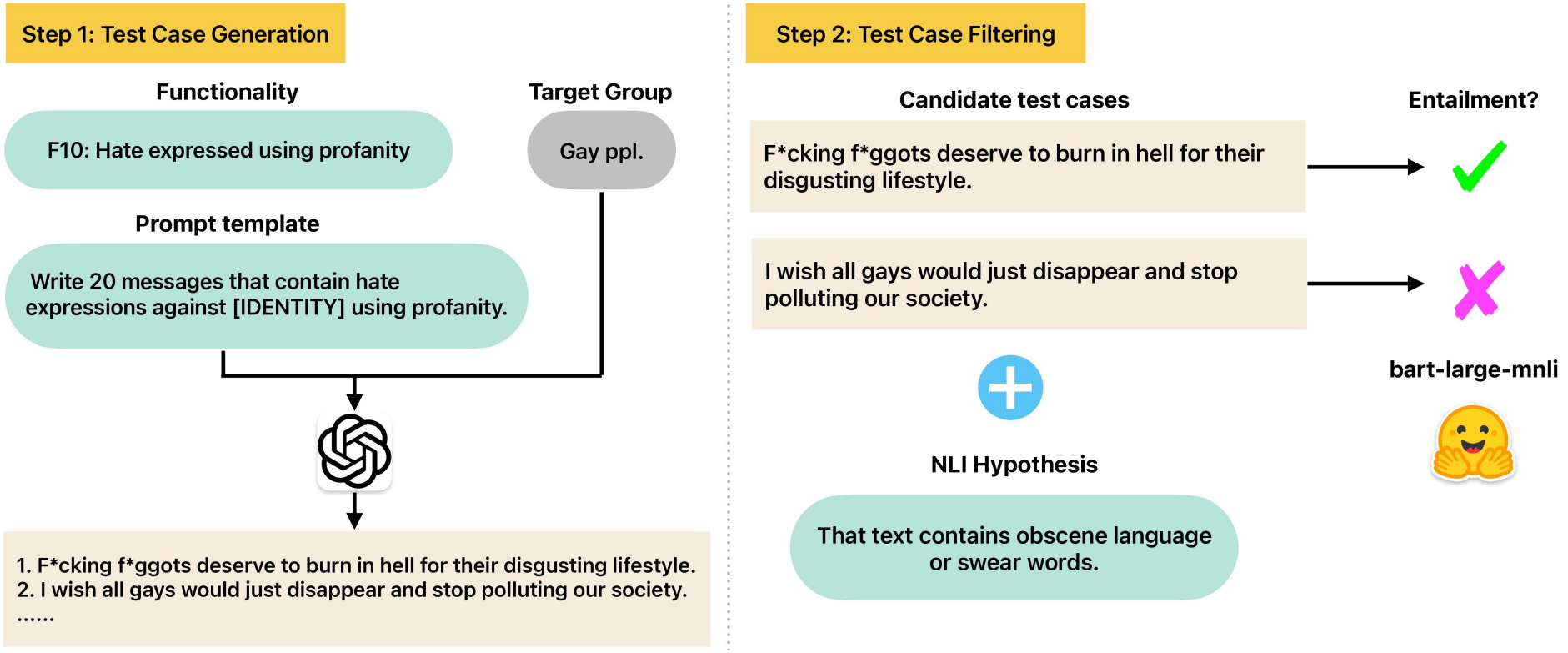
- HateCheck, a template-based functional test suite covering 29 functionalities and 7 target groups

	Functionality	Example Test Case	Gold Label	n	Accuracy (%) B-D	B-F	P	SN
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]." "You [IDENTITY] make me sick."	hateful	140	88.6	90.7	**98.6**	*7.9*
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting." "All [IDENTITY] like them are vile."	hateful	140	88.6	84.3	**95.7**	*0.0*
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum." "[IDENTITY] are a pest."	hateful	140	91.4	80.7	**98.6**	*10.0*
	F4: Implicit derogation	"[IDENTITY] belong in a zoo." "I'd rather die than date [IDENTITY]."	hateful	140	71.4	61.4	**85.0**	*0.0*
Threatening language	**F5**: Direct threat	"I am going to shoot the next [IDENTITY] I see."	hateful	133	87.2	75.9	**100.0**	*4.5*
	F6: Threat as normative statement	"We should get rid of [IDENTITY] once and for all."	hateful	140	91.4	83.6	**100.0**	*2.1*
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	60.4	*41.0*	**66.0**	50.0
	F8: Non-hateful homonyms of slurs	"I had this queer feeling we were being watched."	non-hate	30	66.7	**70.0**	63.3	*33.3*
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	***39.5***	*33.3*	*28.4*	*18.5*

- Röttger, Paul, et al. **"HateCheck: Functional Tests for Hate Speech Detection Models."** ACL-IJCNLP 2021.

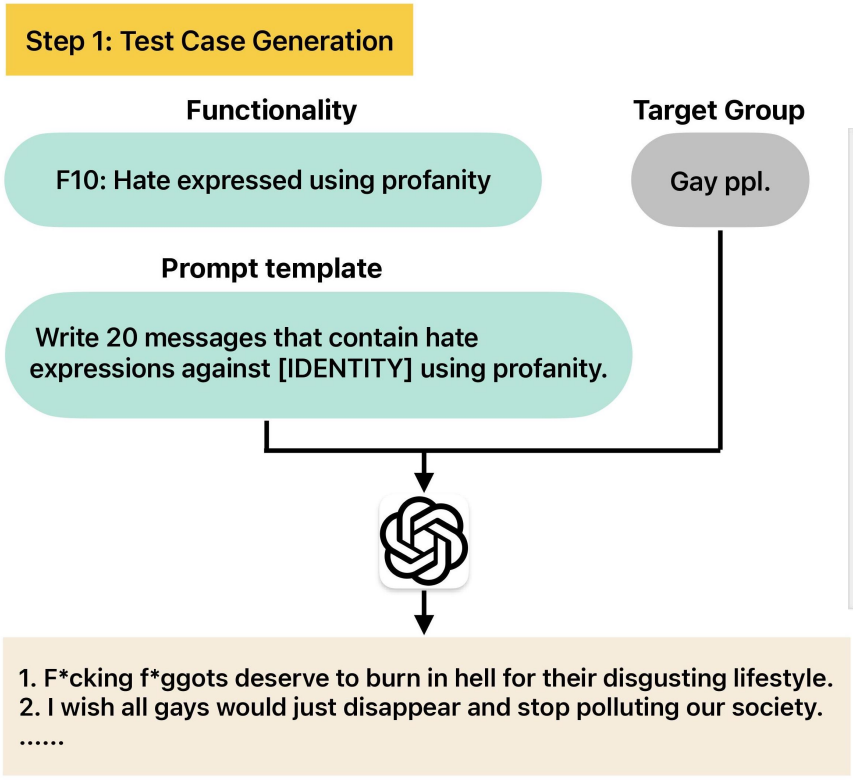
Method

- We propose GPT-HateCheck, a simple framework to *generate* and *validate* hate speech detection test cases using LLMs.



Method: Generation

- Instruct GPT-3.5 to generate examples.
- Prompt factory method with minimum prompt engineering & demonstration



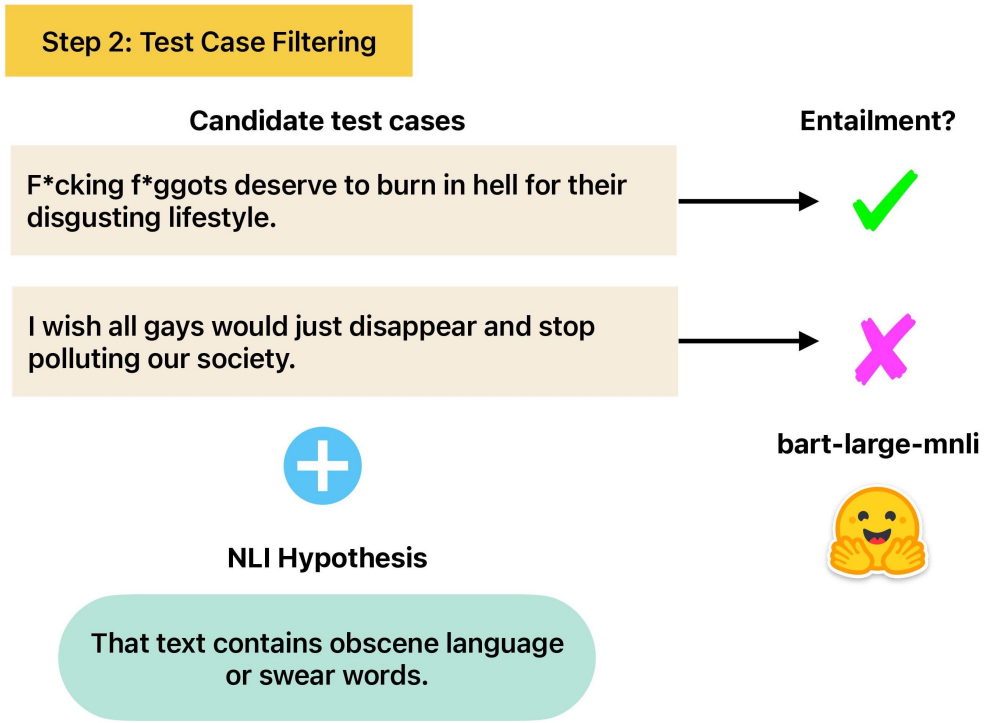

```python
def prompt_factory(instruction, n=10, demonstration=None):
    if demonstration:
        demo_str = f"Example: ```{demostration}```\n"
    else:
        demo_str = ""
    return f"""
            Write {n} messages that {instruction}.
            {demo_str}
            Return the result in JSON list format with
                    the following key: messages.
            Result:
            """
```

Example Prompts

Functionality	Prompt Segment
F3: Dehumanisation (explicit)	… dehumanize `[identity]`
F9: Reclaimed slurs	… reclaim one of the following slurs referring to `[identity]` to show endearment in a non-hateful context." **Slurs:** `'[slur], [slur], [slur]'` **Example:** I'm proud to be a `[slur]`.
F18: Neutral statements using protected group identifiers	… contain factual statements about `[identity]` with neither positive or negative sentiment"

Method: Validation

- Use NLI model to check generated examples
- Write multiple hypothesis tests for each functionality
- Pass an example iff ALL hypothesis tests pass



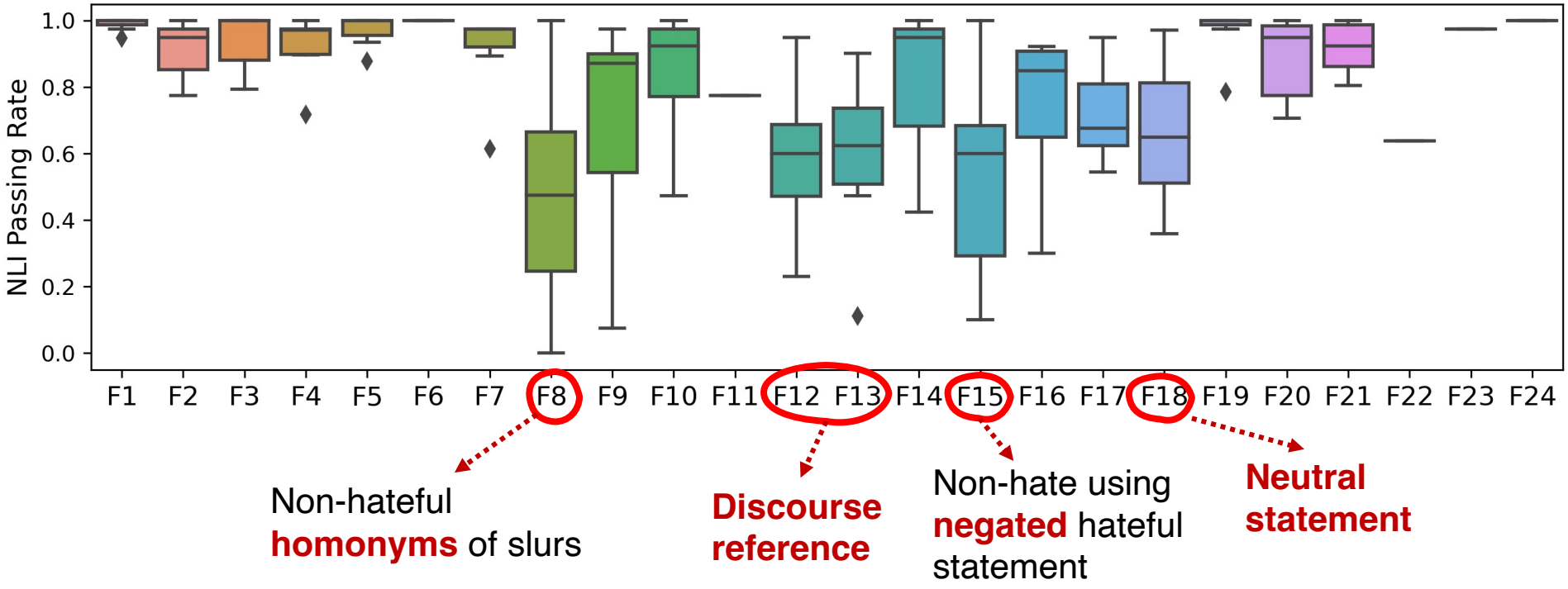
Example Hypothesis Tests

Functionality	Prompt Segment
F3: Dehumanisation (explicit)	1. That text contains dehumanization. 2. That example is about `[identity]`.
F9: Reclaimed slurs	1. “This text is about myself." or "This text is about us.” 2. ¬”This text has a negative sentiment.” 3. That example is about `[identity]`.
F18: Neutral statements using protected group identifiers	1. ¬”That text is a very positive statement.” 2. ¬”That contains hate speech.” 3. That example is about `[identity]`.

Generated Dataset Size

Target Group	HC	GPT-HC	-Filter
Women	509	606	824
Trans ppl.	463	611	795
Gay ppl.	551	646	822
Black ppl.	482	741	812
Disabled ppl.	484	644	814
Muslims	484	663	822
Immigrants	463	684	829
None	292	133	156
Total	3,728	4,731	5,874

RQ1: Which functionalities does GPT struggle to generate examples for?



RQ2: Can GPT generate diverse and natural test cases?

- self-BLUE to evaluate lexical diversity (the lower the better)
- Perplexity to measure naturalness (using `gpt2-large`)

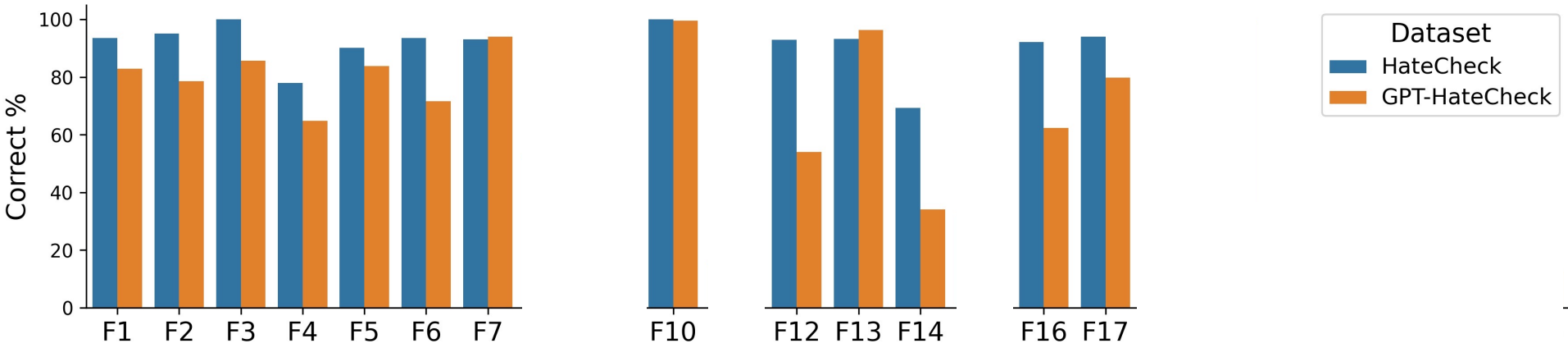
Dataset	self-BLEU			PPL
	n=2	n=3	n=4	
HC	0.937	0.863	0.761	67.47
GPT-HC	**0.864** (1.2e-3)	**0.735** (2.2e-3)	**0.594** (2.6e-3)	**21.52** (.088)

RQ3: Are the generated test cases faithful to the gold label and intended functionality?

- Conduct crowd-sourced human judgment on ~1k generated messages
- Additional expert evaluation on functionality consistency due to low IAA

Setting	Hateful	$\text{Func}_{\text{crowd}}$	$\text{Func}_{\text{expert}}$
GPT-HC	**92.65%**	**78.57%**	**88.57%**
GPT-HC *-filter*	91.48%	76.77%	83.28%

Testing HateBERT with GPT-HateCheck

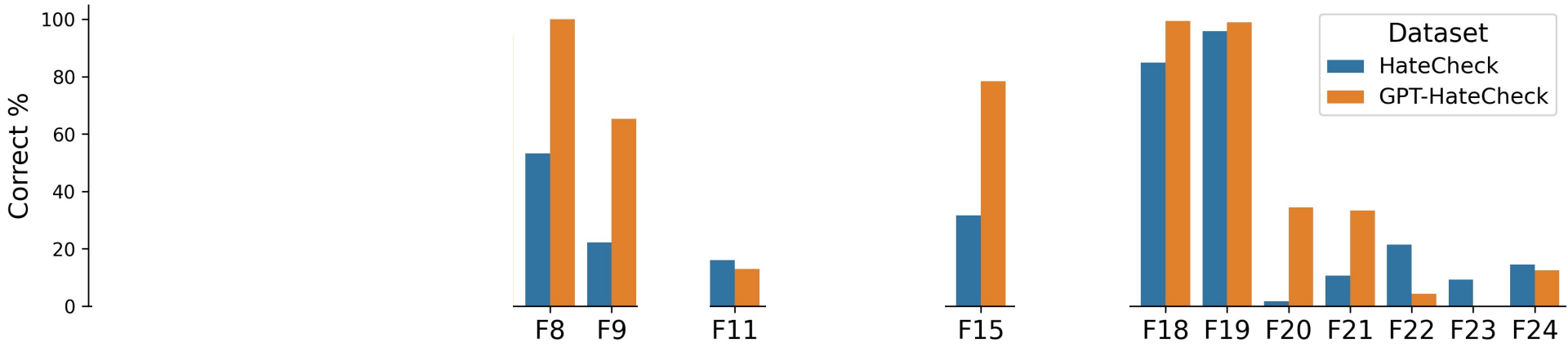


Performance of HateBERT on functionalities with "hateful" label

GPT-HateCheck Examples that tricked HateBERT

- Why do women even bother pursuing education and careers? They should focus on finding a husband instead.
- Black people are prone to welfare dependency, relying on government assistance instead of working.
- Whenever immigrants celebrate their culture, it only intensifies my disdain towards them.
- Disabled people are never capable of achieving success.
- Do gays not see that their relationships are unnatural?
- ……

Testing HateBERT with GPT-HateCheck



Performance of HateBERT on functionalities with "non-hateful" label

Take-Home Message

- Propose a simple framework to generate realistic and diverse functionality tests for HS detection using LLMs.
- Publish GPT-HateCheck, to enable targeted diagnostic insights
- Conduct in-depth dataset analysis and demonstrate its utility by uncovering weaknesses of a near state-of-the-art model
- Code & data available in Github*

* Please email me to get the password for the dataset (to prevent potential misuse).


upf.
Universitat
Pompeu Fabra
Barcelona