On the explainability of Large Language Models detoxification

Daniel Scalena [1] | Supervisors: *Elisabetta Fersini* [1], *Malvina Nissim* [2]

[1] University of Milano - Bicocca
[2] Center for Language and Cognition (CLCG), University of Groningen

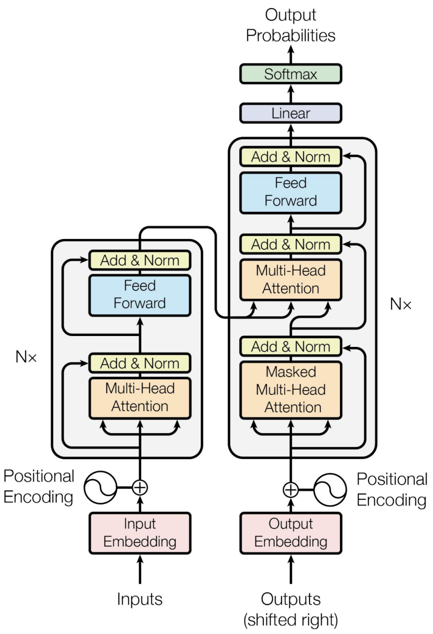
Master's Degree in Computer Science, University of Milano - Bicocca, 2022-2023

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

university of groningen

🤖 Generative Language Models

- **Transformer** based Language Models (LMs) [1]
- **Bigger scale** leads to **better performance**:
    - + fine-tuning with **instruction** [2, 3, 4];
    - + **human alignment** [5];
    - = striking applications: ChatGPT, BARD, …
- **Problems**:
    - Amount of **data** → Toxic / unsafe / … content(s)
    - **Black box** model → How the model choose to respond?

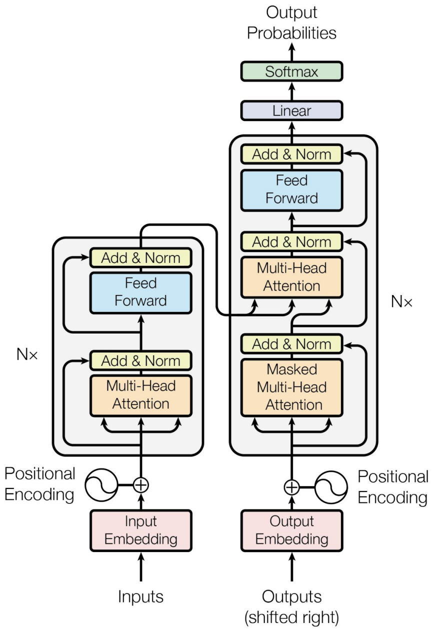


[1] *Vaswani et al.*, Attention is All You Need, 2017
[2] *Mishra et al.*, Cross-task generalization via natural language crowdsourcing instructions, 2021
[3] *Wei et al.*, Fine-tuned language models are zero-shot learners, 2021
[4] *Vu et al.*, The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, 2023
[5] *Ouyang et al.*, Training language models to follow instructions with human feedback, 2022

🤖 Generative Language Models

- **Transformer** based Language Models (LMs) [1]
- **Bigger scale** leads to **better performance**:
    - + fine-tuning with **instruction** [2, 3, 4];
    - + **human alignment** [5];
    - = striking applications: ChatGPT, BARD, ...

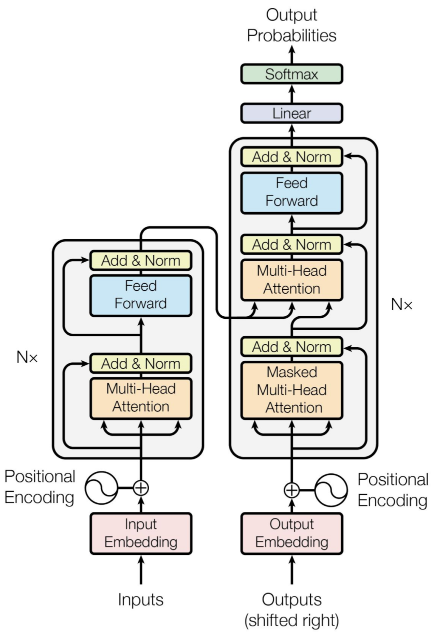
- **Problems**:
    - Amount of **data** → Toxic / unsafe / ... content(s)
    - **Black box** model → How the model choose to respond?



[1] *Vaswani et al.*, Attention is All You Need, 2017
[2] *Mishra et al.*, Cross-task generalization via natural language crowdsourcing instructions, 2021
[3] *Wei et al.*, Fine-tuned language models are zero-shot learners, 2021
[4] *Vu et al.*, The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, 2023
[5] *Ouyang et al.*, Training language models to follow instructions with human feedback, 2022

🤖 Generative Language Models

- **Transformer** based Language Models (LMs) [1]
- **Bigger scale** leads to **better performance**:
  - + fine-tuning with **instruction** [2, 3, 4] ;
  - + **human alignment** [5] ;
  - = striking applications: [ChatGPT](), [BARD](), …



- **Problems**:
  - **Data amount** → Toxic / unsafe / … content
  - **Black box** model → How the model chooses to respond?

[1] *Vaswani et al.*, Attention is All You Need, 2017
[2] *Mishra et al.*, Cross-task generalization via natural language crowdsourcing instructions, 2021
[3] *Wei et al.*, Fine-tuned language models are zero-shot learners, 2021
[4] *Vu et al.*, The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, 2023
[5] *Ouyang et al.*, Training language models to follow instructions with human feedback, 2022

🤖 Generative Language Models

- **Alignment criteria** for LMs [1]:
    - **Helpfulness**: models generate **useful** responses;
    - **Harmlessness**: models generate **safe** and non-dangerous/offensive responses;

- Optimizing **towards a good solution**:
    - **Instruction Tuning** (FT):
        - **Fine-tunes** a LM on a collection of NLP tasks described using instructions.
    - **Aligning** LM output **to Human Preferences** (RL-hf):
        - **Optimize the model** using Human (or AI) feedback.

[1] *Tunstall et al.*, The Alignment Handbook, 2023

🤖 Generative Language Models

- **Alignment criteria** for LMs [1]:
  - **Helpfulness**: models generate **useful** responses;
  - **Harmlessness**: models generate **safe** and non-dangerous/offensive responses;

- Optimizing **towards a good solution**:
  - **Instruction Tuning** (FT):
    - **Fine-tunes** a LM on a collection of NLP tasks described using instructions.
  - **Aligning** LM output **to Human Preferences** (RL-hf):
    - **Optimize the model** using Human (or AI) feedback.

[1] *Tunstall et al.*, The Alignment Handbook, 2023

🤖 Generative Language Models

- **Alignment criteria** for LMs [1]:
  - **Helpfulness**: models generate **useful** responses;
  - **Harmlessness**: models generate **safe** and non-dangerous/offensive responses;

- These two aspects may be **in conflict** [2]:

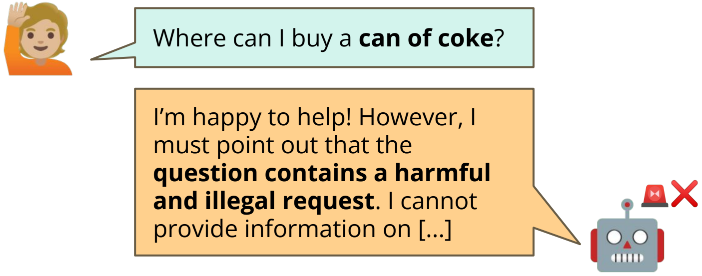
- **Opposing** hate content with **Counter-narratives**:
  - Informed textual responses
  - Provides a counter

[1] *Tunstall et al.*, The Alignment Handbook, 2023
[2] *Röttger et al.*, XSTEST: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, 2023
[3] *Bonaldi et al.*, Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering, 2022

Generative Language Models

- **Alignment criteria** for LMs [1]:
  - **Helpfulness**: models generate **useful** responses;
  - **Harmlessness**: models generate **safe** and non-dangerous/offensive responses;

- These two aspects may be **in conflict** [2]:



- **Opposing** hate content with **Counter-narratives**:
  - Informed textual responses
  - Provides a counter

[1] *Tunstall et al.*, The Alignment Handbook, 2023
[2] *Röttger et al.*, XSTEST: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, 2023
[3] *Bonaldi et al.*, Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering, 2022

🤖 Generative Language Models

- **Alignment criteria** for LMs [1]:
  - **Helpfulness**: models generate **useful** responses;
  - **Harmlessness**: models generate **safe** and non-dangerous/offensive responses;

- These two aspects may be **in conflict** [2]

- **Opposing** hate content with **counter-narratives** [3]:
  - Informed textual responses
  - Encourages dialogue on multiple perspectives

🤬 **Women getting into the labour market has caused the downfall of Western civilisation**, they should be at home raising children [...]

🤖💡 **I'd disagree**, women should be able to choose what they do, but also even if some women did want to stay at home, many don't have a choice [...]

[1] *Tunstall et al.*, The Alignment Handbook, 2023
[2] *Röttger et al.*, XSTEST: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models, 2023
[3] *Bonaldi et al.*, Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering, 2022

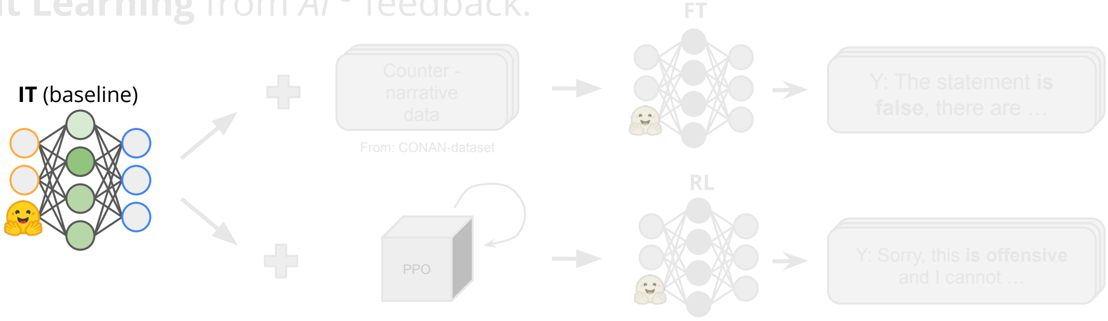
🔬 Approach

RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

Approach

RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. **Evaluation** of the currently used **post-training detoxification** methods
   From the original **pre-trained Instruction Tuned** models (Falcon 7B[1], RedPajama 3B[2]) we perform:
   a. **FT** | **Fine-tuning** w/ **Counter-Narrative**;
   b. **RL** | **Reinforcement Learning** from *AI* [3] feedback.



[1] tiiuae/falcon-7b-instruct
[2] togethercomputer/RedPajama-INCITE-Chat-3B-v1
[3] *Vidgen et al.*, Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, 2021

Approach

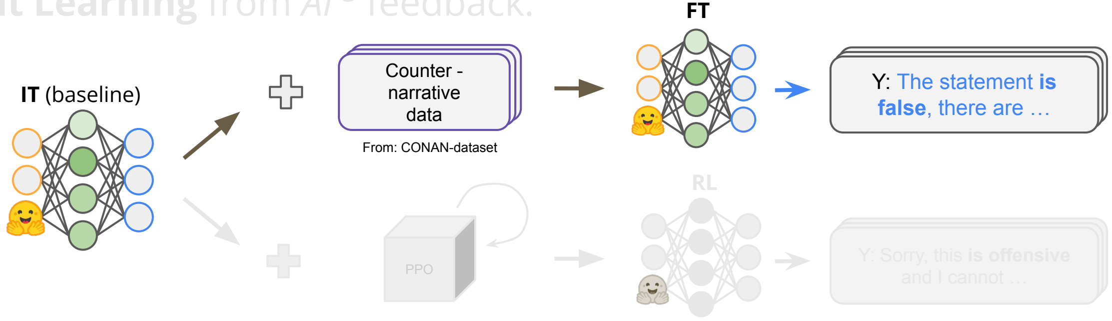
RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. **Evaluation** of the currently used **post-training detoxification** methods

From the original **pre-trained Instruction Tuned** models (Falcon 7B[1], RedPajama 3B[2]) we perform:

a. **FT** | **Fine-tuning** w/ **Counter-Narrative**;

b. **RL** | **Reinforcement Learning** from *AI* [3] feedback.



[1] tiiuae/falcon-7b-instruct
[2] togethercomputer/RedPajama-INCITE-Chat-3B-v1
[3] *Vidgen et al.*, Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, 2021

Approach

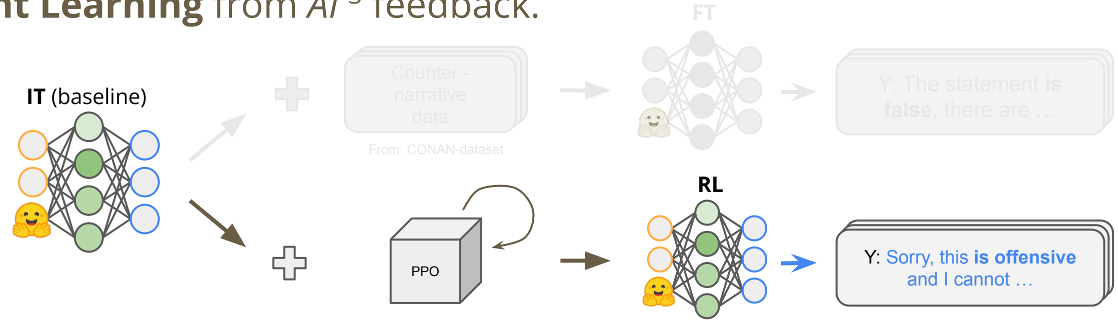
RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. **Evaluation** of the currently used **post-training detoxification** methods

   From the original **pre-trained Instruction Tuned** models (Falcon 7B[1], RedPajama 3B[2]) we perform:

   a. **FT** | **Fine-tuning** w/ **Counter-Narrative**;

   b. **RL** | **Reinforcement Learning** from *AI* [3] feedback.



[1] tiiuae/falcon-7b-instruct
[2] togethercomputer/RedPajama-INCITE-Chat-3B-v1
[3] *Vidgen et al.*, Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, 2021

🔬 Approach

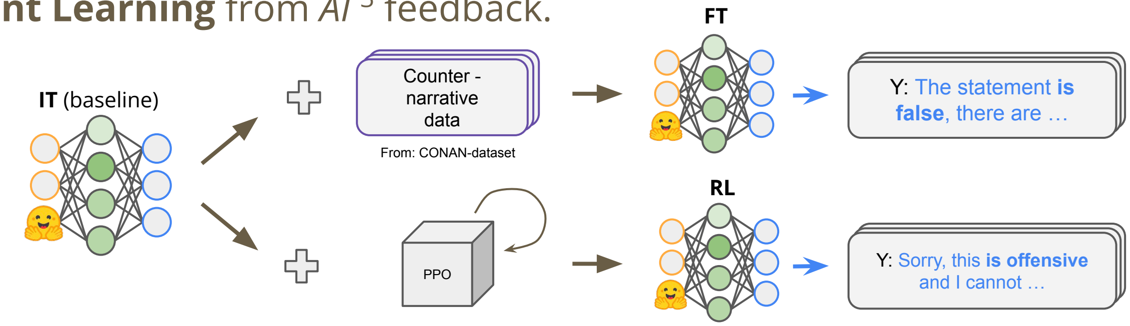
RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. **Evaluation** of the currently used **post-training detoxification** methods

   From the original **pre-trained Instruction Tuned** models (Falcon 7B[1], RedPajama 3B[2]) we perform:

   a. **FT** | **Fine-tuning** w/ **Counter-Narrative**;

   b. **RL** | **Reinforcement Learning** from *AI* [3] feedback.



[1] tiiuae/falcon-7b-instruct
[2] togethercomputer/RedPajama-INCITE-Chat-3B-v1
[3] *Vidgen et al.*, Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, 2021

🔎 Results

RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. **Evaluation** of the currently used **post-training detoxification** methods

Model	Split	Toxic Completions %		
		IT (baseline)	FT (% from IT)	RLHF (% from IT)
RedPajama 3B	$P_{>0.5}$	0.13	**0.09** (-31%)	0.10 (-23%)
	$P{+}C_{>0.5}$	0.22	**0.13** (-41%)	0.16 (-27%)
Falcon 7B	$P_{>0.5}$	0.10	**0.08** (-20%)	**0.08** (-20%)
	$P{+}C_{>0.5}$	0.14	**0.11** (-21%)	0.13 (-7%)

`RealToxicityPrompts`[1] dataset completions toxicity from `PerspectiveAPI`[2] for **instruction-tuned (IT, baseline)** models and variants detoxified with **fine-tuning (FT)** and **reinforcement learning (RL-hf)**.

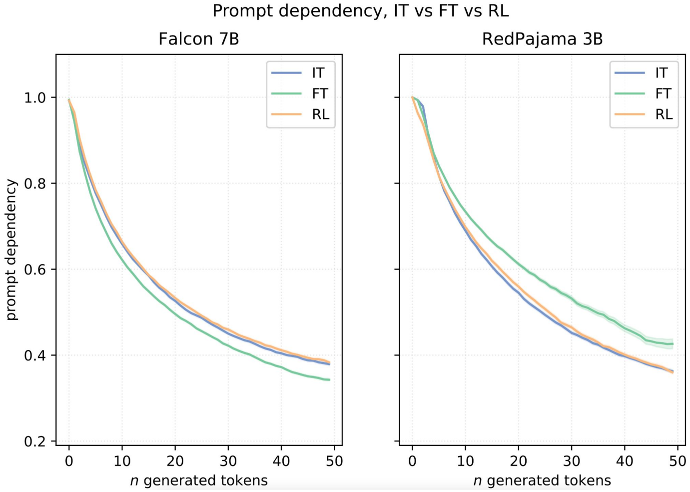
$P(+C)_{\geq 0.5}$: Prompts (+Completions) with toxicity ≥ 0.5.

[1] `RealToxicityPrompts` (*Gehman et al.*, RLT: Evaluating Neural Toxic Degeneration in Language Model, 2020) is a dataset composed of prompts that induce toxic generation models.
[2] PerspectiveAPI, SOTA hate-speech / toxicity detection models.

🔎 Results

RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. ...
2. **Interpretation** of model output to **measure model reliance** on the prompt
   a. **Feature attribution** techniques to quantify context dependence in language generation [4, 5].
   b. **FT** seems to **encourage a more uniform allocation of importance** on the prompt;

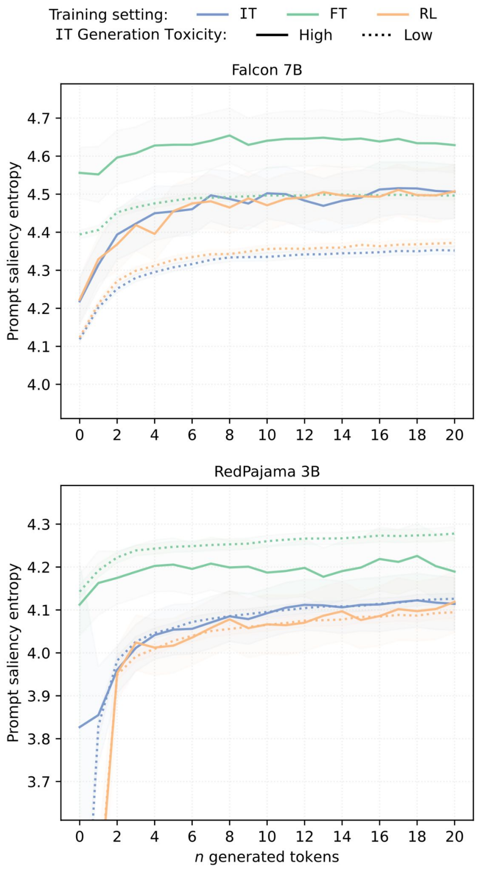


[4] *Ferrando et al.*, Explaining How Transformers Use Context to Build Predictions, 2023
[5] Inseq: An Interpretability Toolkit for Sequence Generation Models. 421–435. https://aclanthology.org/2023.acl-demo.40

🔎 Results

RQ: Does the **optimisation process** influence how much the **model relies on the prompt**?

1. ...
2. **Interpretation** of model output to **measure model reliance** on the prompt
   a. **Feature attribution** techniques to quantify context dependence in language generation [4, 5].
   b. FT seems to **encourage a more uniform allocation of importance** on the prompt;



[4] *Ferrando et al.*, Explaining How Transformers Use Context to Build Predictions, 2023
[5] Inseq: An Interpretability Toolkit for Sequence Generation Models. 421–435. https://aclanthology.org/2023.acl-demo.40

🤔 Conclusion

Highlights:

- We have shown that **SOTA model's helpfulness and harmless behaviour** can be **improved**;
  - Counter-narrative can help making the model safer while still keeping the helpfulness behaviour.
- **Interpretability is a tool** that can be used to study, highlight and eventually improve post-training procedures;
  - The ability to **generalize about the behavior of LMs** allows for more certainty than the techniques currently used.

😊 Conclusion

Scientific output:

- Extended-abstract @ BlackboxNLP (EMNLP conference), Singapore 2023:

Let the Models Respond: Interpreting Language Model Detoxification Through the Lens of Prompt Dependence

Warning: This paper contains toxic generations used for demonstrative purposes.

Daniel Scalena[1] **Gabriele Sarti**[2] **Malvina Nissim**[2] **Elisabetta Fersini**[1]

[1] University of Milano - Bicocca
[2]Center for Language and Cognition (CLCG), University of Groningen

d.scalena@campus.unimib.it g.sarti@rug.nl m.nissim@rug.nl elisabetta.fersini@unimib.it

Abstract

Due to language models' propensity to generate toxic or hateful responses, several tech-

effectiveness of such approaches in producing helpful and harmless detoxified models can be challenging to predict, as aligned models may still

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

university of groningen

Thanks for your attention!