On Assessing the Political Biases of Multilingual Large Language Models

Paul Lerner, Laurène Cave, Hal Daumé III, Léo Labat, Gaël Lejeune, Pierre-Antoine Lequeu, Benjamin Piwowarski, Nazanin Shafiabadi, François Yvon

29th of June 2025 – EALM@TALN 2025, Marseille

Sorbonne Université, CNRS, ISIR
lerner@isir.upmc.fr

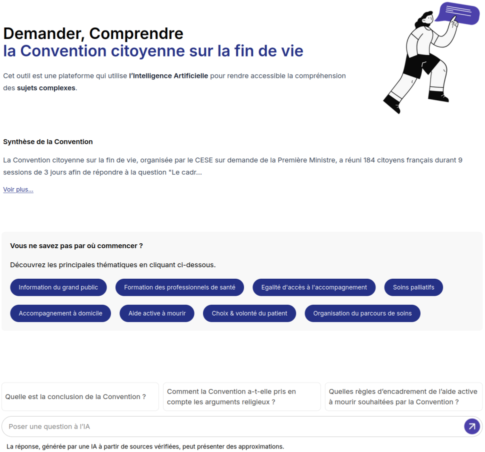
Introduction

Benefits and Risks of LLMs for Democratic Deliberation

- 100M+ people interact with LLMs everyday through ChatGPT et al.

Benefits and Risks of LLMs for Democratic Deliberation

- 100M+ people interact with LLMs everyday through ChatGPT et al.
- LLMs are used to foster democratic participation (make.org)

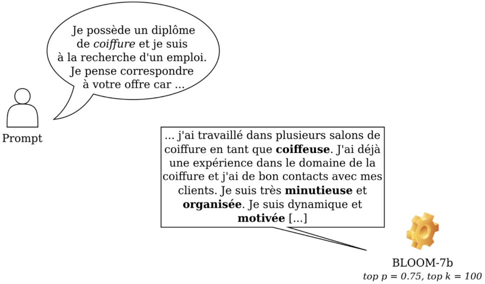


Benefits and Risks of LLMs for Democratic Deliberation

- 100M+ people interact with LLMs everyday through ChatGPT et al.
- LLMs are used to foster democratic participation (make.org)
- LLMs are used to complement polling results (fairgen.ai)

Benefits and Risks of LLMs for Democratic Deliberation

- 100M+ people interact with LLMs everyday through ChatGPT et al.
- LLMs are used to foster democratic participation (make.org)
- LLMs are used to complement polling results (fairgen.ai)
- LLMs generate racist, sexist, and “toxic” texts



Ducel et al. (2024)

Benefits and Risks of LLMs for Democratic Deliberation

- 100M+ people interact with LLMs everyday through ChatGPT et al.
- LLMs are used to foster democratic participation (make.org)
- LLMs are used to complement polling results (fairgen.ai)
- LLMs generate racist, sexist, and “toxic” texts
- **What are the political biases of LLMs? How can we assess them?**

Benefits and Risks of LLMs for Democratic Deliberation

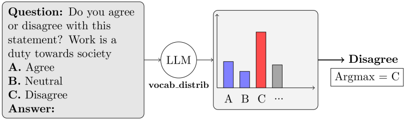
- 100M+ people interact with LLMs everyday through ChatGPT et al.
- LLMs are used to foster democratic participation (make.org)
- LLMs are used to complement polling results (fairgen.ai)
- LLMs generate racist, sexist, and “toxic” texts
- **What are the political biases of LLMs? How can we assess them?**

⚠️ position paper ⚠️

⚠️ no results ⚠️

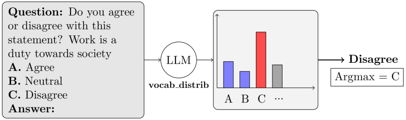
Can we ask an LLM what it thinks?

- Administer questionnaires that give left-right scores (e.g. politicalcompass.org)

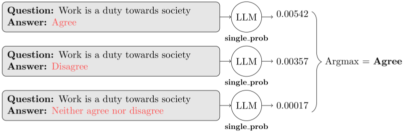


Can we ask an LLM what it thinks?

- Administer questionnaires that give left-right scores (e.g. politicalcompass.org)

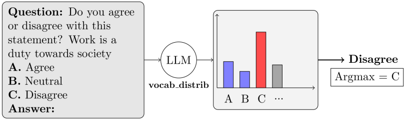


- But how do you get **the** answer?



Can we ask an LLM what it thinks?

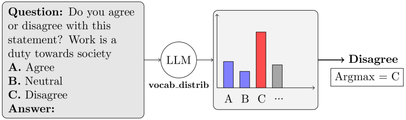
- Administer questionnaires that give left-right scores (e.g. politicalcompass.org)



- But how do you get **the** answer?
- Should there be a space between every option? Or `"\n"`? Or `"\t"`?

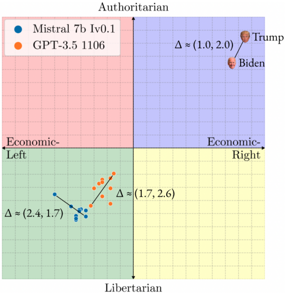
Can we ask an LLM what it thinks?

- Administer questionnaires that give left-right scores (e.g. politicalcompass.org)



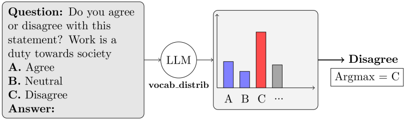
- But how do you get **the** answer?
- Should there be a space between every option? Or `"\n"`? Or `"\t"`?

- Makes a big difference (Boelaert et al., 2024; Ceron et al., 2024; Röttger et al., 2024)



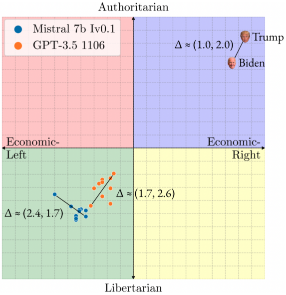
Can we ask an LLM what it thinks?

- Administer questionnaires that give left-right scores (e.g. politicalcompass.org)



- But how do you get **the** answer?
- Should there be a space between every option? Or `"\n"`? Or `"\t"`?

- Makes a big difference (Boelaert et al., 2024; Ceron et al., 2024; Röttger et al., 2024)



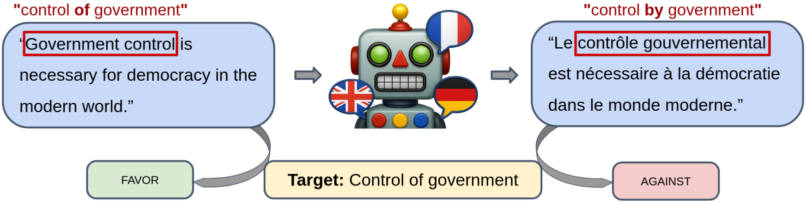
- left-right scoring per question is arbitrary

Constraining the Setting

- Assessing biases of LLMs for
  - machine translation
  - writing assistance
  - summarization

Constraining the Setting

- Assessing biases of LLMs for
  - machine translation
  - writing assistance
  - summarization



Constraining the Setting

- Assessing biases of LLMs for
  - machine translation
  - writing assistance
  - summarization

Sortir de l'euro

Souveraineté européenne

Inflation

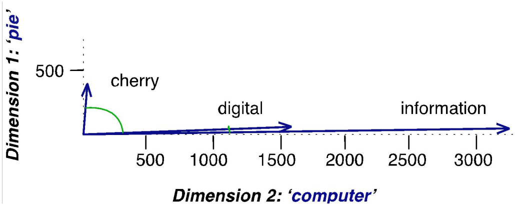
Inflation. Sortir de l'euro.

Embedding Politics

Word Embedding 101: the Distributional Hypothesis

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

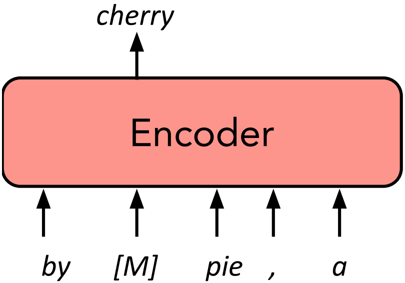
	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...



If two words appear in similar contexts, they are synonyms (Harris, 1954)

Word Embedding 101: Masked Language Modeling

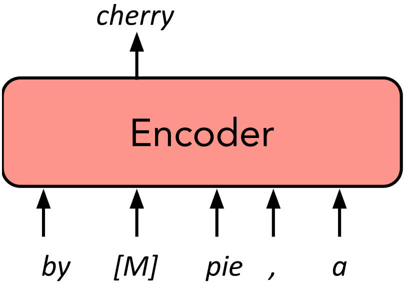
is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet



Devlin et al. (2019)

Word Embedding 101: Masked Language Modeling

is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet



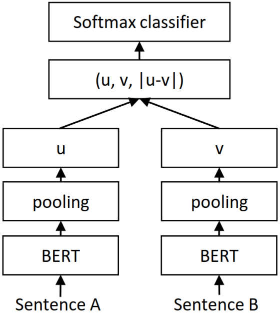
Devlin et al. (2019)

- One vector per token → What about higher-level embeddings?

Sentence Embedding 101: NLI/STS

Met my first girlfriend that way. ✗ I didn't meet my first girlfriend until later.

At 8:34, the Boston Center controller received a third transmission from American 11 ✓ The Boston Center controller got a third transmission from American 11.

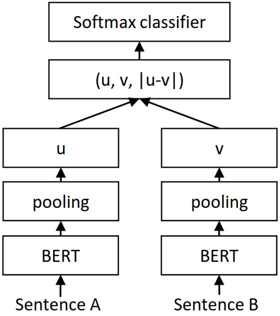


Reimers and Gurevych (2019)

Sentence Embedding 101: NLI/STS

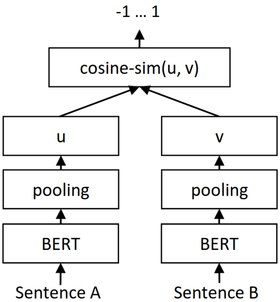
Met my first girlfriend that way. ✗ I didn't meet my first girlfriend until later.

At 8:34, the Boston Center controller received a third transmission from American 11 ✓ The Boston Center controller got a third transmission from American 11.



Reimers and Gurevych (2019)

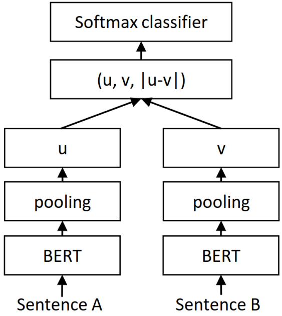
(5) The two sentences are completely equivalent, as they mean the same thing.
The bird is bathing in the sink.
Birdie is washing itself in the water basin.
(4) The two sentences are mostly equivalent, but some unimportant details differ.
In May 2010, the troops attempted to invade Kabul.
The US army invaded Kabul on May 7th last year, 2010.
(3) The two sentences are roughly equivalent, but some important information differs/missing.
John said he is considered a witness but not a suspect.
"He is not a suspect anymore." John said.
(2) The two sentences are not equivalent, but share some details.
They flew out of the nest in groups.
They flew into the nest together.
(1) The two sentences are not equivalent, but are on the same topic.
The woman is playing the violin.
The young lady enjoys listening to the guitar.
(0) The two sentences are on different topics.
John went horse back riding at dawn with a whole group of friends.
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.



Sentence Embedding 101: NLI/STS

Met my first girlfriend that way. ✗ I didn't meet my first girlfriend until later.

At 8:34, the Boston Center controller received a third transmission from American 11 ✓ The Boston Center controller got a third transmission from American 11.

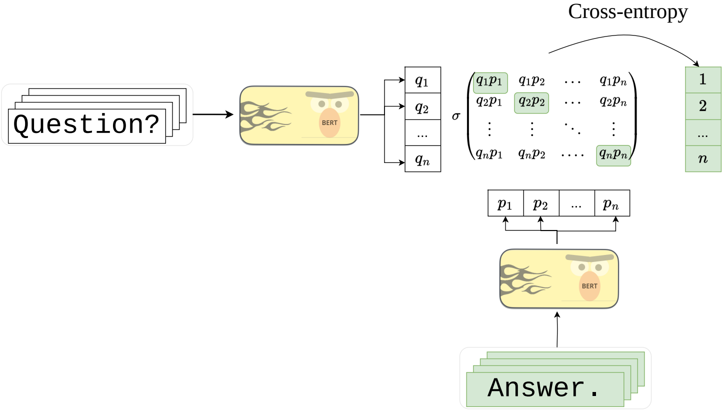


Reimers and Gurevych (2019)

(5) The two sentences are completely equivalent, as they mean the same thing.
The bird is bathing in the sink.
Birdie is washing itself in the water basin.
(4) The two sentences are mostly equivalent, but some unimportant details differ.
In May 2010, the troops attempted to invade Kabul.
The US army invaded Kabul on May 7th last year, 2010.
(3) The two sentences are roughly equivalent, but some important information differs/missing.
John said he is considered a witness but not a suspect.
"He is not a suspect anymore." John said.
(2) The two sentences are not equivalent, but share some details.
They flew out of the nest in groups.
They flew into the nest together.
(1) The two sentences are not equivalent, but are on the same topic.
The woman is playing the violin.
The young lady enjoys listening to the guitar.
(0) The two sentences are on different topics.
John went horse back riding at dawn with a whole group of friends.
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

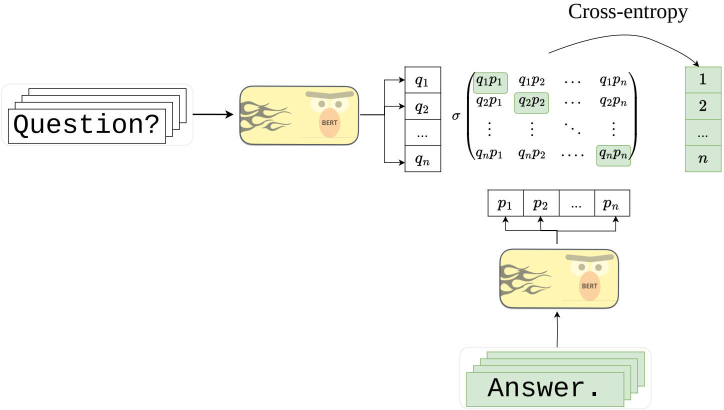
- expensive annotation $\Longrightarrow$ limited to English

Sentence Embedding 101: query-document pairs



Karpukhin et al. (2020); Lee et al. (2019); Xiong et al. (2021)

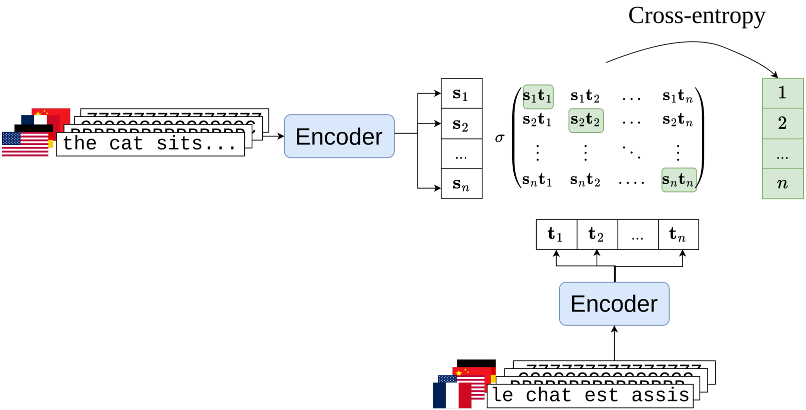
Sentence Embedding 101: query-document pairs



Karpukhin et al. (2020); Lee et al. (2019); Xiong et al. (2021)

- expensive annotation $\implies$ limited to English

Sentence Embedding 101: cross-lingual embeddings



Artetxe and Schwenk (2019); Feng et al. (2022)

Semantic similarity? What about politics?

- existing methods give “semantic” representation

Semantic similarity? What about politics?

- existing methods give "semantic" representation
- $\implies$ "**Liberty is an essential part of democracy**"
  $\approx$ "**Liberty is** not **an essential part of democracy**"
- $\implies$ "**Liberty** is an essential part **of** democracy"
  $\neq$ "Democracies should always guarantee the **liberty of** their citizen"

Training an embedding that models political opinions

What the model should learn:

- **topic-stance**, e.g. Manifesto (Merz et al., 2016): 3,219 programs of 954 parties over 78 years and 60 countries in 40 languages

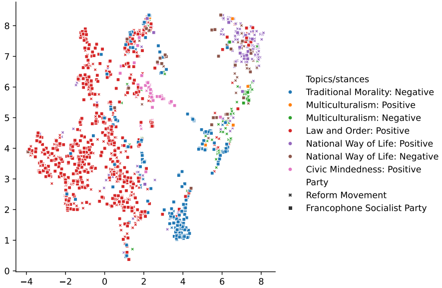
Britain is struggling to emerge from a long and difficult recession/Families 305
are finding it hard to make ends meet/Millions are unemployed, and millions 603
more have taken pay cuts or reduced hours to stay in their jobs/And there 701
are deeper problems too/Britain, for all its many strengths, is still too unequal 503
and unfair, a country where the circumstances of your birth and the income of
your parents still profoundly affect your chances in life/Our children's future is 503
threatened by climate change, which we have done far too little to stop/And 501
the political system is in crisis./ 305

Britain needs a fresh start/We need hope for a different, better future./ 305 305

Training an embedding that models political opinions

What the model should learn:

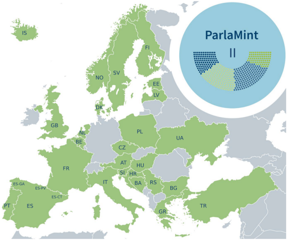
- **topic-stance**, e.g. Manifesto (Merz et al., 2016): 3,219 programs of 954 parties over 78 years and 60 countries in 40 languages



Training an embedding that models political opinions

What the model should learn:

- **topic-stance**, e.g. Manifesto (Merz et al., 2016): 3,219 programs of 954 parties over 78 years and 60 countries in 40 languages
- **party of the speaker**, e.g. Parlamint (Erjavec et al., 2024): parliamentary debates of 29 countries in 31 languages over 28 years



Erjavec et al. (2024)

Training an embedding that models political opinions

What the model should learn:

- **topic-stance**, e.g. Manifesto (Merz et al., 2016): 3,219 programs of 954 parties over 78 years and 60 countries in 40 languages
- **party of the speaker**, e.g. Parlamint (Erjavec et al., 2024): parliamentary debates of 29 countries in 31 languages over 28 years
- **source of newspaper** (e.g. custom dataset through scraping)


MARDI 27 MAI 2025
AFFAIRE BASTIEN VIVÈS Enquête sur un procès polémique
Libération
GAZA INJUSTIFIABLE
Refusant toujours de laisser entrer l'aide humanitaire, l'armée israélienne poursuit ses bombardements aveugles. La plupart de ses alliés ne l'acceptent plus.
L'APPEL DES 300 ÉCRIVAINS
«Nous ne pouvons plus nous contenter du mot "horreur", il faut nommer le "génocide"»


LE FIGARO
FORMULE 1
SCIENCES PO
La stratégie de conquête des Frères musulmans
Le chinois BYD s'attaque au marché des petites voitures électriques en Europe
Après l'appel entre Trump et Poutine, l'Europe et l'Ukraine redoutent un retrait américain
Que de temps perdu !

Training an embedding that models political opinions

What the model should learn:

- **topic-stance**, e.g. Manifesto (Merz et al., 2016): 3,219 programs of 954 parties over 78 years and 60 countries in 40 languages
- **party of the speaker**, e.g. Parlamint (Erjavec et al., 2024): parliamentary debates of 29 countries in 31 languages over 28 years
- **source of newspaper** (e.g. custom dataset through scraping)

- $\implies$ train a multi-task classifier

Constraining a multilingual representation space

- Multilingual classifier might learn one subspace per language

Constraining a multilingual representation space

- Multilingual classifier might learn one subspace per language
- Enforce common multilingual space using bi-parallel data (NLLB: 2,656 language pairs, 450GB)

Constraining a multilingual representation space

- Multilingual classifier might learn one subspace per language
- Enforce common multilingual space using bi-parallel data (NLLB: 2,656 language pairs, 450GB)
- Evaluate: Precision@1 on multi-parallel EuroParl (21 languages $\times$ 23,647 sentences)

Constraining a multilingual representation space

- Multilingual classifier might learn one subspace per language
- Enforce common multilingual space using bi-parallel data (NLLB: 2,656 language pairs, 450GB)

- Evaluate: Precision@1 on multi-parallel EuroParl (21 languages $\times$ 23,647 sentences)
- Fine-tune from XLM-RoBERTa

Model	**P@1**
LaBSE	93.4
MEXMA	89.1
Bi-Encoder (ours)	90.9
Manifesto Classifier (unconstrained)	46.7
Classifier (constrained)	**79.7**

Meta-evaluation: probing source of newspaper article

- Linear probing to evaluate multiple embeddings
- Custom dataset of 12 French newspapers
- Year 2024, temporal split: 4 months for train/dev/test (50K+ articles each)
- In addition to multilingual constraint: continual MLM training using CC-100

Meta-evaluation: probing source of newspaper article

- Linear probing to evaluate multiple embeddings
- Custom dataset of 12 French newspapers
- Year 2024, temporal split: 4 months for train/dev/test (50K+ articles each)
- In addition to multilingual constraint: continual MLM training using CC-100

Model	Accuracy
majority	17.6
LaBSE	53.1
MEXMA	57.1
Bi-Encoder (ours)	65.1
Manifesto Classifier (unconstrained)	57.9
Parlamint Classifier (unconstrained)	62.5
Classifier (unconstrained)	62.8
Classifier + MLM	67.4
Classifier (constrained)	**70.1**

Assessing Biases

Summarization biases: more formally

document $D = (s_1, s_2, ..., s_N)$

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0, 1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

Summarization biases: more formally

document $D = (s_1, s_2, ..., s_N)$

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0, 1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

word embeddings $\mathbf{h}_n = \text{encoder}(s_n)$, $\mathbf{h} \in \mathbb{R}^{L \times d}$

sentence embedding $\mathbf{s}_n = \text{pool}(\mathbf{h}_n) \in \mathbb{R}^d$

Summarization biases: more formally

document $D = (s_1, s_2, ..., s_N)$

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0,1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

word embeddings $\mathbf{h}_n = \text{encoder}(s_n)$, $\mathbf{h} \in \mathbb{R}^{L \times d}$

sentence embedding $\mathbf{s}_n = \text{pool}(\mathbf{h}_n) \in \mathbb{R}^d$

clusters $c = (c_1, c_2, ..., c_N)$, $c_n = \text{cluster}(\mathbf{s}_n)$, $c_n \in \{0,1\}^K$ (one-hot)

ideological distribution $p = (p_1, p_2, ..., p_K)$, $p_k = \frac{\sum_{n=1}^{N} c_{nk}}{N}$,
$p_k \in [0,1]^K, \sum_{k=1}^{K} p_k = 1$

Summarization biases: more formally

document $D = (s_1, s_2, ..., s_N)$

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0,1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

word embeddings $\mathbf{h}_n = \text{encoder}(s_n)$, $\mathbf{h} \in \mathbb{R}^{L \times d}$

sentence embedding $\mathbf{s}_n = \text{pool}(\mathbf{h}_n) \in \mathbb{R}^d$

clusters $c = (c_1, c_2, ..., c_N)$, $c_n = \text{cluster}(\mathbf{s}_n)$, $c_n \in \{0,1\}^K$ (one-hot)

ideological distribution $p = (p_1, p_2, ..., p_K)$, $p_k = \frac{\sum_{n=1}^{N} c_{nk}}{N}$,
$p_k \in [0,1]^K, \sum_{k=1}^{K} p_k = 1$

summarization bias $b = \text{KL}(p, q) = \sum_{k=1}^{K} p_k \log\left(\frac{p_k}{q_k}\right)$, q = ideological distribution of summary, $b \geq 0$

- small b: the two distributions match well, small bias
- great b: the two distributions do not match well, great bias

Machine Translation biases: more formally

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0, 1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

word embeddings $\mathbf{h}_n = \text{encoder}(s_n)$, $\mathbf{h} \in \mathbb{R}^{L \times d}$

sentence embedding $\mathbf{s}_n = \text{pool}(\mathbf{h}_n) \in \mathbb{R}^d$

Machine Translation biases: more formally

sentence $s_n = (w_1, w_2, ..., w_L)$

word $w_l \in \{0,1\}^V$ (one-hot: $\sum_{l=1}^{L} w_l = 1$)

word embeddings $\mathbf{h}_n = \text{encoder}(s_n)$, $\mathbf{h} \in \mathbb{R}^{L \times d}$

sentence embedding $\mathbf{s}_n = \text{pool}(\mathbf{h}_n) \in \mathbb{R}^d$

clusters $c = (c_1, c_2, ..., c_N)$, $c_n = \text{cluster}(\mathbf{s}_n)$, $c_n \in \{0,1\}^K$

bias $b = 1 - \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} c_{nk} c'_{nk}}{N}$, $b \in [0,1]$, c' = clusters of the translation

Conclusion

- Assessing the political biases of LLMs is a timely matter
- Existing methods rely on questionnaires which is brittle
- We propose a method for embedding political text

Conclusion

- Assessing the political biases of LLMs is a timely matter
- Existing methods rely on questionnaires which is brittle
- We propose a method for embedding political text
- Stay tuned for results

Thank you for your attention!

Paul Lerner, Laurène Cave, Hal Daumé III, Léo Labat, Gaël Lejeune, Pierre-Antoine Lequeu, Benjamin Piwowarski, Nazanin Shafiabadi, François Yvon

29th of June 2025 – EALM@TALN 2025, Marseille

Sorbonne Université, CNRS, ISIR
lerner@isir.upmc.fr

Mikel Artetxe and Holger Schwenk. 2019. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Julien Boelaert, Samuel Coavoux, Etienne Ollion, Ivaylo D. Petev, and Patrick Präg. 2024. How do Generative Language Models Answer Opinion Polls?

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. "You'll be a nurse, my son!" Automatically assessing gender biases in autoregressive language models in French and Italian. *Language Resources and Evaluation*.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, and et al. 2024. ParlaMint II: Advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2-3):146–162.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Nicolas Merz, Sven Regel, and Jirka Lewandowski. 2016. The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics*, 3(2):2053168016643346.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance.

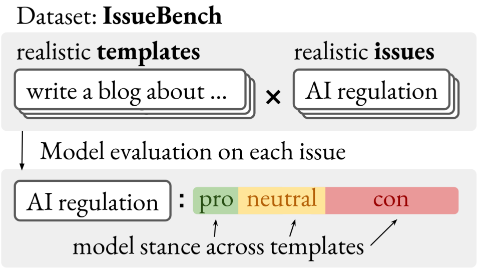
Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3):148.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Related Work



Röttger et al. (2025)

- Limited to English language/US politics
- Uses an LLM for stance detection: possible meta-bias?
- More constrained than questionnaires but still finds little coherence among LLMs outputs, needs to filter
- Pro-neutral-con stance framework rigid: what does it mean to be pro domestic violence? pro holocaust?