Comment mesurer les biais politiques des grands modèles de langue multilingues?

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

Comment mesurer les biais politiques des grands modèles de langue multilingues?

Paul Lerner[1] Laurène Cave[2] Hal Daumé III[3] Léo Labat[1] Gaël Lejeune[2]
Pierre-Antoine Lequeu[1] Benjamin Piwowarski[1] Nazanin Shafiabadi[1]
François Yvon[1]
(1) Sorbonne Université, CNRS, ISIR, 75005, Paris, France
(2) Sorbonne Université, STIH/CERES, 75006, Paris, France
(3) University of Maryland, College Park, USA
<prenom>.<nom>@sorbonne-universite.fr

RÉSUMÉ

Nous proposons une nouvelle méthode pour mesurer les biais politiques des grands modèles de langue multilingues pour la traduction automatique, l'aide à la rédaction et le résumé automatique. Nous nous appuyons sur une représentation dense des opinions politiques exprimées dans les textes, apprise de façon faiblement supervisée.

ABSTRACT

On Assessing the Political Biases of Multilingual Large Language Models

We propose a new method to assess the political biases of Multilingual Large Language Models for machine translation, writing assistance, and summarization. We rely on a weakly supervised dense representation of political opinions expressed in texts.

MOTS-CLÉS : biais politique, grand modèle de langue, plongement de phrase.

KEYWORDS: Political Bias, Large Language Model, Sentence Embedding.

ARTICLE : **Contribution originelle.**

1 Introduction

Large Language Models (LLMs) are now ubiquitous in Natural Language Processing (NLP) and are used to address most tasks conceptualized so far (Zhao *et al.*, 2023). Moreover, their reach has extended far beyond NLP academics : chatbot-based LLMs, such as ChatGPT, are used daily by 100M+ people, naturally spiking the interest of sociologists (Bail, 2024). Indeed, as LLMs have been found to accurately simulate parts of human behaviors (Park *et al.*, 2023), several sociologists argue that LLMs could partly complement traditional polling methods (Argyle *et al.*, 2023) or even deliberation (Small *et al.*, 2023). Businesses are already proposing such SaaS solutions for polling. [1] On the other hand, LLMs have been found to encode or even amplify biases present in their training data, leading to racist, sexist, or other forms of "toxic" generations (Nangia *et al.*, 2020; Cheng

1. E.g., https://www.fairgen.ai/, https://www.expectedparrot.com/ and https://www.syntheticusers.com/.

et al., 2023; Lum *et al.*, 2024; Ducel *et al.*, 2024; Gallegos *et al.*, 2024). This hints at the necessity to examine the political biases encoded in LLMs and the potential risks they pose. Most studies addressing this question have analogized LLMs to individual citizens, evaluating their responses to opinion polls, questionnaires, or values surveys (Feng *et al.*, 2023; Rozado, 2023; Hartmann *et al.*, 2023; Santurkar *et al.*, 2023; Durmus *et al.*, 2024; Motoki *et al.*, 2024). Since then, several papers have demonstrated the brittleness of such approaches, as LLMs are sensitive to details of the prompt format unnoticeable to humans (e.g., using a new line instead of a space) and also lack internal coherence (Boelaert *et al.*, 2024; Röttger *et al.*, 2024; Ceron *et al.*, 2024). We argue that the political biases of LLMs can be better measured in controlled settings, such as summarization, translation, or writing assistance. For example, when summarizing a debate between a Democrat and a Republican, an unbiased LLM should not outweigh one side over the other but accurately summarize each – possibly conflicting – opinion expressed in the debate. To automatically assess how well LLMs can generate such balanced summaries, we propose to rely on dedicated dense representations of political opinions expressed in texts and introduce a methodology for learning such embedding spaces.

Politics is intimately linked to language and culture. Our work more broadly aims to assess the linguistic and cultural biases encoded in LLMs. For example, Wendler *et al.* (2024) show that English-centric multilingual LLMs exhibit internal representations biased towards English. We can therefore expect their political biases to lean towards a predominantly US-centric perspective, as suggested by the results of Durmus *et al.* (2024) (notwithstanding the methodological issues discussed above). We aim to design a multilingual representation space in order to assess the multilingual biases of LLMs for machine translation (e.g., an opinion expressed in English should be preserved when translated to French). In doing so, we go beyond Röttger *et al.* (2025), who only assess the bias of LLMs for writing assistance in English.

2 Learning a Multilingual Representation of Political Opinions

Word embeddings rely on Harris' Distributional Hypothesis which states that words sharing similar contexts have related meanings (because they are substitutable ; Harris, 1954). Word embeddings implement that idea by either reducing the dimension of a word co-occurrence matrix (Deerwester *et al.*, 1990; Levy & Goldberg, 2014), predicting the context of a target word (Mikolov *et al.*, 2013), or, conversely, the target word from its context (Masked Language Modeling or MLM for short ; Mikolov *et al.*, 2013; Devlin *et al.*, 2019). Building upon this research and on MLM-pretrained models, sentence embeddings are trained either by : (i) regressing the manually annotated semantic similarity of a pair of sentences (often limited to English because of the annotation cost ; Reimers & Gurevych, 2019) ; (ii) contrasting semantically related texts such as query-document pairs, either manually annotated (Lee *et al.*, 2019; Karpukhin *et al.*, 2020; Xiong *et al.*, 2021), automatically generated (Lewis *et al.*, 2021), or using various heuristics (Ram *et al.*, 2022) (also often limited to English) ; (iii) contrasting translations from parallel corpora (Artetxe & Schwenk, 2019; Feng *et al.*, 2022; Janeiro *et al.*, 2024) ; (iv) any combination of the above (Chen *et al.*, 2024).

However, these methods lead to vaguely defined "semantic" representations of texts that fail to capture political opinions or stances. For example, "*Liberty is an essential part of democracy*" would likely be more similar to "*Liberty is <u>not</u> an essential part of democracy*" rather than "*Democracies should always guarantee the liberty of their citizen*" (see Kletz *et al.* (2023) for studies related to negation). To fill this gap, we leverage both press articles as well as annotated corpora primarily designed for

the comparative analysis of political stances to train a multi-task classifier : (i) the Manifesto corpus (Merz *et al.*, 2016), an effort of the V-DEM project, covers 3,219 political programs of 954 political parties over 78 years and 60 countries in 40 languages, annotated with a fine-grained topic-stance (e.g., "Military : positive") ; (ii) Parlamint (Erjavec *et al.*, 2024) contains parliamentary debates of 29 countries in 31 languages over 28 years, annotated with the party of the speaker ; (iii) we collect our own dataset of article paired with its publishing newspaper through web scraping. [2] To also constrain our model to share its representation across languages, we add a cross-lingual contrastive objective similar to LaBSE (Feng *et al.*, 2022), using available parallel corpora. Last but not least, to reduce the domain shift when processing texts from other domains and avoid catastrophic forgetting, we continue the self-supervised pretraining of our model using MLM. To assess the quality of the resulting embedding space, we turn to linear probing (Alain & Bengio, 2017; Liu *et al.*, 2019) using stance datasets such as X-Stance (Vamvas & Sennrich, 2020). [3]

Our preliminary results suggest that training a multilingual classifier on topic-stance only leads to a moderate linguistic bias that contrasting translations from parallel corpora can mitigate.

3 Assessing the Political Biases of LLMs

With an accurate representation of political opinions, we can now assess the political biases of LLMs. We focus on three tasks : translation, writing assistance, and summarization. For translation, we argue that an unbiased translator should preserve in the target language the opinion expressed in the source ; consequently, both representations should be close. We quantify this by clustering a set of texts and their automatic translation. If the source text does not cluster with its translation, we deem the translation to be biased. We experiment with X-Stance. Likewise, for writing assistance, we argue that the machine-generated text should preserve the opinion expressed in the original text and apply the same method. The case is more complex for summarization as we expect multiple – possibly conflicting – opinions to be expressed in the source documents. Therefore, we argue that only the *distribution* of opinions expressed in the summary should match that of the source documents. To do so, we compute the distribution of opinions by segmenting the text into sentences and clustering sentences over the entire dataset. We quantify the match of the two distributions using Kullback–Leibler (KL) divergence. An ideally unbiased summary would perfectly match the distribution of the source documents and obtain a KL divergence of 0.

4 Conclusion

We propose a new method for assessing the political biases of Multilingual Large Language Models, relying on a dense representation of political opinions expressed in texts. This representation is learned through weak supervision combined with parallel corpora to constrain a shared representation across languages. In future work, we will experiment to demonstrate that our method leads to an accurate assessment of biases, unlike the state-of-the-art questionnaires methods.

2. This objective is similar to Liu *et al.* (2022) but does not require story-level alignment of articles nor access to the political leaning of the newspaper and is not restricted to English.

3. This analysis resembles the work of Kim *et al.* (2025), who find that LLMs such as Llama-2 (Touvron *et al.*, 2023) encode the political stance of several politicians and newspapers in some attention heads. However, their work is more related to interpretability rather than bias assessment.

Acknowledgments

We thank the anonymous reviewers for their knowledgeable comments.

This research was funded by Bpifrance under the project AI For Democracy - Democratic Commons, one of seven winners of Bpifrance's 'Digital Commons for Generative AI' call for projects, conducted as part of the France 2030 investment plan.

Références

ALAIN G. & BENGIO Y. (2017). Understanding intermediate layers using linear classifier probes.

ARGYLE L. P., BUSBY E. C., FULDA N., GUBLER J. R., RYTTING C. & WINGATE D. (2023). Out of One, Many : Using Language Models to Simulate Human Samples. *Political Analysis*, **31**(3), 337–351. DOI : 10.1017/pan.2023.2.

ARTETXE M. & SCHWENK H. (2019). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3197–3203, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1309.

BAIL C. A. (2024). Can generative AI improve social science ? *Proceedings of the National Academy of Sciences*, **121**(21), e2314021121. DOI : 10.1073/pnas.2314021121.

BOELAERT J., COAVOUX S., OLLION E., PETEV I. D. & PRÄG P. (2024). How do Generative Language Models Answer Opinion Polls ? DOI : 10.31235/osf.io/r2pnb.

CERON T., FALK N., BARIĆ A., NIKOLAEV D. & PADÓ S. (2024). Beyond Prompt Brittleness : Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, **12**, 1378–1400. DOI : 10.1162/tacl_a_00710.

CHEN J., XIAO S., ZHANG P., LUO K., LIAN D. & LIU Z. (2024). M3-Embedding : Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 2318–2335, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.137.

CHENG M., DURMUS E. & JURAFSKY D. (2023). Marked Personas : Using Natural Language Prompts to Measure Stereotypes in Language Models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1504–1532, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.84.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DUCEL F., NÉVÉOL A. & FORT K. (2024). "You'll be a nurse, my son !" Automatically assessing gender biases in autoregressive language models in French and Italian. *Language Resources and Evaluation*. DOI : 10.1007/s10579-024-09780-6.

DURMUS E., NGUYEN K., LIAO T., SCHIEFER N., ASKELL A., BAKHTIN A., CHEN C., HATFIELD-DODDS Z., HERNANDEZ D., JOSEPH N., LOVITT L., MCCANDLISH S., SIKDER O., TAMKIN A., THAMKUL J., KAPLAN J., CLARK J. & GANGULI D. (2024). Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

ERJAVEC T., KOPP M., LJUBEŠIĆ N., KUZMAN T., RAYSON P., OSENOVA P., OGRODNICZUK M., ÇÖLTEKIN Ç., KORŽINEK D., MEDEN K., SKUBIC J., RUPNIK P., AGNOLONI T., AIRES J., BARKARSON S., BARTOLINI R., BEL N., CALZADA PÉREZ M., DARGIS R., DIWERSY S., GAVRIILIDOU M., VAN HEUSDEN R., IRUSKIETA M., KAHUSK N., KRYVENKO A., LIGETI-NAGY N., MAGARIÑOS C., MÖLDER M., NAVARRETTA C., SIMOV K., TUNGLAND L. M., TUOMINEN J., VIDLER J., VLADU A. I., WISSIK T., YRJÄNÄINEN V. & FIŠER D. (2024). ParlaMint II : Advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*. DOI : 10.1007/s10579-024-09798-w.

FENG F., YANG Y., CER D., ARIVAZHAGAN N. & WANG W. (2022). Language-agnostic BERT Sentence Embedding. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 878–891, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.62.

FENG S., PARK C. Y., LIU Y. & TSVETKOV Y. (2023). From pretraining data to language models to downstream tasks : Tracking the trails of political biases leading to unfair NLP models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11737–11762, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.656.

GALLEGOS I. O., ROSSI R. A., BARROW J., TANJIM M. M., KIM S., DERNONCOURT F., YU T., ZHANG R. & AHMED N. K. (2024). Bias and Fairness in Large Language Models : A Survey. *Computational Linguistics*, **50**(3), 1097–1179. DOI : 10.1162/coli_a_00524.

HARRIS Z. S. (1954). Distributional Structure. *WORD*, **10**(2-3), 146–162. DOI : 10.1080/00437956.1954.11659520.

HARTMANN J., SCHWENZOW J. & WITTE M. (2023). The political ideology of conversational AI : Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. DOI : 10.2139/ssrn.4316084.

JANEIRO J. M., PIWOWARSKI B., GALLINARI P. & BARRAULT L. (2024). MEXMA : Token-level objectives improve sentence representations.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781.

KIM J., EVANS J. & SCHEIN A. (2025). Linear representations of political perspective emerge in large language models. In *The Thirteenth International Conference on Learning Representations*.

KLETZ D., AMSILI P. & CANDITO M. (2023). The self-contained negation test set. In Y. BELINKOV, S. HAO, J. JUMELET, N. KIM, A. MCCARTHY & H. MOHEBBI, Éds., *Proceedings of the 6th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*, p. 212–221, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.blackboxnlp-1.16.

LEE K., CHANG M.-W. & TOUTANOVA K. (2019). Latent retrieval for weakly supervised open domain question answering. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6086–6096, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1612.

LEVY O. & GOLDBERG Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 27 : Curran Associates, Inc.

LEWIS P., WU Y., LIU L., MINERVINI P., KÜTTLER H., PIKTUS A., STENETORP P. & RIEDEL S. (2021). PAQ : 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, **9**, 1098–1115. DOI : 10.1162/tacl_a_00415.

LIU N. F., GARDNER M., BELINKOV Y., PETERS M. E. & SMITH N. A. (2019). Linguistic knowledge and transferability of contextual representations. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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)*, p. 1073–1094, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1112.

LIU Y., ZHANG X. F., WEGSMAN D., BEAUCHAMP N. & WANG L. (2022). POLITICS : Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 1354–1374, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.101.

LUM K., ANTHIS J. R., NAGPAL C. & D'AMOUR A. (2024). Bias in Language Models : Beyond Trick Tests and Toward RUTEd Evaluation. DOI : 10.48550/arXiv.2402.12649.

MERZ N., REGEL S. & LEWANDOWSKI J. (2016). The Manifesto Corpus : A new resource for research on political parties and quantitative text analysis. *Research & Politics*, **3**(2), 2053168016643346. DOI : 10.1177/2053168016643346.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient Estimation of Word Representations in Vector Space. DOI : 10.48550/arXiv.1301.3781.

MOTOKI F., PINHO NETO V. & RODRIGUES V. (2024). More human than human : Measuring ChatGPT political bias. *Public Choice*, **198**(1), 3–23. DOI : 10.1007/s11127-023-01097-2.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). CrowS-Pairs : A Challenge Dataset for Measuring Social Biases in Masked Language Models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1953–1967, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

PARK J. S., O'BRIEN J., CAI C. J., MORRIS M. R., LIANG P. & BERNSTEIN M. S. (2023). Generative agents : Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3586183.3606763.

RAM O., SHACHAF G., LEVY O., BERANT J. & GLOBERSON A. (2022). Learning to Retrieve Passages without Supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 2687–2700, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.193.

REIMERS N. & GUREVYCH I. (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), p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

RÖTTGER P., HINCK M., HOFMANN V., HACKENBURG K., PYATKIN V., BRAHMAN F. & HOVY D. (2025). IssueBench : Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance. DOI : 10.48550/arXiv.2502.08395.

RÖTTGER P., HOFMANN V., PYATKIN V., HINCK M., KIRK H. R., SCHÜTZE H. & HOVY D. (2024). Political Compass or Spinning Arrow ? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models.

ROZADO D. (2023). The Political Biases of ChatGPT. *Social Sciences*, **12**(3), 148. DOI : 10.3390/socsci12030148.

SANTURKAR S., DURMUS E., LADHAK F., LEE C., LIANG P. & HASHIMOTO T. (2023). Whose Opinions Do Language Models Reflect ? In *Proceedings of the 40th International Conference on Machine Learning*, p. 29971–30004 : PMLR.

SMALL C. T., VENDROV I., DURMUS E., HOMAEI H., BARRY E., CORNEBISE J., SUZMAN T., GANGULI D. & MEGILL C. (2023). Opportunities and risks of llms for scalable deliberation with polis. DOI : 10.48550/arXiv.2306.11932.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open Foundation and Fine-Tuned Chat Models.

VAMVAS J. & SENNRICH R. (2020). X-Stance : A Multilingual Multi-Target Dataset for Stance Detection. DOI : 10.48550/arXiv.2003.08385.

WENDLER C., VESELOVSKY V., MONEA G. & WEST R. (2024). Do Llamas Work in English ? On the Latent Language of Multilingual Transformers. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15366–15394, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.820.

XIONG L., XIONG C., LI Y., TANG K.-F., LIU J., BENNETT P. N., AHMED J. & OVERWIJK A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

ZHAO W. X., ZHOU K., LI J., TANG T., WANG X., HOU Y., MIN Y., ZHANG B., ZHANG J., DONG Z., DU Y., YANG C., CHEN Y., CHEN Z., JIANG J., REN R., LI Y., TANG X., LIU Z., LIU P., NIE J.-Y. & WEN J.-R. (2023). A Survey of Large Language Models. DOI : 10.48550/arXiv.2303.18223.