27. Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., de Rijke, M.: Query performance prediction using relevance judgments generated by large language models. arXiv preprint arXiv:2404.01012 (2024)
28. Meta, L.T.A..: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
29. Mu, Y., Wu, B., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., Song, X.: Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. ArXiv **abs/2305.14310** (2023). https://doi.org/10.48550/arXiv.2305.14310
30. Murr, L., Grainger, M., Gao, D.: Testing llms on code generation with varying levels of prompt specificity. ArXiv **abs/2311.07599** (2023). https://doi.org/10.48550/arXiv.2311.07599
31. Poesina, E., Costache, A.V., Chifu, A.G., Mothe, J., Ionescu, R.T.: Pqpp: A joint benchmark for text-to-image prompt and query performance prediction (2024), https://arxiv.org/abs/2406.04746
32. Raj, H., Gupta, V., Rosati, D., Majumdar, S.: Semantic consistency for assuring reliability of large language models. ArXiv **abs/2308.09138** (2023). https://doi.org/10.48550/arXiv.2308.09138
33. Rajapakse, T.C., Yates, A., de Rijke, M.: Simple transformers: Open-source for all. In: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. p. 209–215. SIGIR-AP 2024, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3673791.3698412, https://doi.org/10.1145/3673791.3698412
34. Salamat, S., Arabzadeh, N., Seyedsalehi, S., Bigdeli, A., Zihayat, M., Bagheri, E.: Neural disentanglement of query difficulty and semantics. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 4264–4268 (2023)
35. Saleminezhad, A., Arabzadeh, N., Beheshti, S., Bagheri, E.: Context-aware query term difficulty estimation for performance prediction. In: European Conference on Information Retrieval. pp. 30–39. Springer (2024). https://doi.org/10.1007/978-3-031-56066-8_4
36. Sclar, M., Choi, Y., Tsvetkov, Y., Suhr, A.: Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. arXiv preprint arXiv:2310.11324 (2023)
37. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
38. Yan, Z.: Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). eugeneyan.com (Aug 2024), https://eugeneyan.com/writing/llm-evaluators/
39. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1259, https://aclanthology.org/D18-1259
40. Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Gong, N., Zhang, Y., Xie, X.: Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. ArXiv **abs/2306.04528** (2023). https://doi.org/10.48550/arXiv.2306.04528
41. Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., Chen, K.: Prosa: Assessing and understanding the prompt sensitivity of llms (2024), https://arxiv.org/abs/2410.12405