Towards Truly Open, Language-Specific, Safe, Factual, and Specialized Large Language Models

Preslav Nakov

Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Building 1B, 3rd Floor
Abu Dhabi, UAE

Abstract

First, we will argue for the need for fully transparent open-source large language models (LLMs), and we will describe the efforts of MBZUAI's Institute on Foundation Models (IFM) towards that based on the LLM360 initiative. Second, we will argue for the need for language-specific LLMs, and we will share our experience from building Jais, the world's leading open Arabic-centric foundation and instruction-tuned large language model, Nanda, our recently released open Hindi LLM, and some other models. Third, we will argue for the need for safe LLMs, and we will present Do-Not-Answer, a dataset for evaluating the guardrails of LLMs, which is at the core of the safety mechanisms of our LLMs. Forth, we will argue for the need for factual LLMs, we will discuss the factuality challenges that LLMs pose. We will then present some recent relevant tools for addressing these challenges developed at MBZUAI: (i) OpenFactCheck, a framework for fact-checking LLM output, for building customized fact-checking systems, and for benchmarking LLMs for factuality, (ii) LM-Polygraph, a tool for predicting an LLM's uncertainty in its output using cheap and fast uncertainty quantification techniques, and (iii) LLM-DetectAIve, a tool for machine-generated text detection. Finally, we will argue for the need for specialized models, and we will present the zoo of LLMs currently being developed at MBZUAI's IFM.

Bio:

Bio

Preslav Nakov is Professor and Department Chair for NLP at the Mohamed bin Zayed University of Artificial Intelligence. He is part of the core team at MBZUAI's Institute for Foundation Models that developed Jais, the world's best open-source Arabic-centric LLM, Nanda, the world's best Hindi model, and LLM360, the first truly open LLM. Previously, he was Principal Scientist at the Qatar Computing Research Institute, HBKU, where he led the Tanbih mega-project, developed in collaboration with MIT, which aims to limit the impact of "fake news", propaganda and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking. He received his PhD degree in Computer Science from the University of California at Berkeley, supported by a Fulbright grant. He is Chair-Elect of the European Chapter of the Association for Computational Linguistics (EACL), Secretary of ACL SIGSLAV, and Secretary of the Truth and Trust Online board of trustees. Formerly, he was PC chair of ACL 2022, and President of ACL SIGLEX. He is also member of the editorial board of several journals including Computational Linguistics, TACL, ACM TOIS, IEEE TASL, IEEE TAC, CS&L, NLE, AI Communications, and Frontiers in AI. He authored a Morgan & Claypool book on Semantic Relations between Nominals, two books on computer algorithms, and 250+ research papers. He received a Best Paper Award at ACM WebSci'2022, a Best Long Paper Award at CIKM'2020, a Best Resource Paper Award at EACL'2024, a Best Demo Paper Award (Honorable Mention) at ACL'2020, a Best Task Paper Award (Honorable Mention) at SemEval'2020, a Best Poster Award at SocInfo'2019, and the Young Researcher Award at RANLP'2011. He was also the first to receive the Bulgarian President's John Atanasoff award, named after the inventor of the first automatic electronic digital computer. His research was featured by over 100 news outlets, including Reuters, Forbes, Financial Times, CNN, Boston Globe, Aljazeera, DefenseOne, Business Insider, MIT Technology Review, Science Daily, Popular Science, Fast Company, The Register, WIRED, and Engadget, among others.