SmileyLlama: Modifying Large Language Models for Directed Chemical Space Exploration

Jospeh M. Cavanagh
Kenneth S. Pitzer Theory Center and Department of Chemistry,
University of California, Berkeley, Berkeley, CA, 94720 USA
jmcavanagh@berkeley.edu

Kunyang Sun
Kenneth S. Pitzer Theory Center and Department of Chemistry,
University of California, Berkeley, Berkeley, CA, 94720 USA

Andrew Gritsevskiy
Department of Computer Science, University of Wisconsin–Madison,
Madison, WI, 53706 USA

Dorian Bagni
Kenneth S. Pitzer Theory Center and Department of Chemistry,
University of California, Berkeley, Berkeley, CA, 94720 USA

Thomas D. Bannister
Department of Molecular Medicine, The Herbert Wertheim UF Scripps
Institute for Biomedical Innovation and Technology,
130 Scripps Way, Jupiter, FL, 33458

Teresa Head-Gordon
Kenneth S. Pitzer Theory Center and Department of Chemistry,
University of California, Berkeley, CA, 94720 USA
Departments of Bioengineering and Chemical and Biomolecular Engineering,
University of California, Berkeley, CA, 94720 USA
Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720 USA
thg@berkeley.edu

Abstract

Here we show that a Large Language Model (LLM) can serve as a foundation model for a Chemical Language Model (CLM) which performs at or above the level of CLMs trained solely on chemical SMILES string data. Using supervised fine-tuning (SFT) and direct preference optimization (DPO) on the open-source Llama LLM, we demonstrate that we can train an LLM to respond to prompts such as generating molecules with properties of interest to drug development. This overall framework allows an LLM to not just be a chatbot client for chemistry and materials tasks, but can be adapted to speak more directly as a CLM which can generate molecules with user-specified properties.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).