universität innsbruck

Apollo

Ancient Languages, Open Data, and Language Model Training

by Anna Dolganov & David Smith

Thursday, 25 June 2026 - 6pm
Innrain 52a, Ágnes-Heller-Haus, Seminar room 7

We present the first results from the Apollo project to train LLMs on Ancient Greek and evaluate them on tasks relevant to scholars of the ancient world. Apollo's training data is the largest collection of open-access Greek known to us. Most material dates from before the traditional end of Ancient Greek in 1453, but we also include data in the katharevousa register until 1900. A key part of expanding the source base beyond existing collections is the use of automated transcription. We also developed methods to handle the linguistic variation in our documents without over-normalization. Using this corpus, Apollo was specifically trained to restore missing text surrounded by known context, a task frequently encountered in the study of papyri, inscriptions, and manuscripts. In the final part of the talk, we present some experiments with the model illustrating its capabilities, as well as prospects for future development and expansion to other ancient corpora.