

HKUST ECON Seminar

19 May, 2026, 10:00 am – 12:00 nn

Menu Pricing of Large Language Models

Prof Alex Smolin, Toulouse School of Economic

(with Dirk Bergemann and Alessandro Bonatti)

Abstract:

We develop a framework for the optimal pricing and product design of LLMs in which a provider sells menus of token budgets to users who differ in their valuations across a continuum of tasks. Under a homogeneous production technology, we show that users' high-dimensional type profiles are summarized by a scalar index, reducing the seller's problem to one-dimensional screening. The optimal mechanism takes the form of committed-spend contracts: buyers pay for a budget that they allocate across token classes priced at marginal cost. We extend the analysis to environments with multiple differentiated models and to competition between a proprietary leader and an open-source fringe, showing that competitive pressure reshapes both the intensive and extensive margins of compute provision. Each element of our theory (token-budget menus, maximum- and minimum-spend plans, multi-model versioning, and linear API pricing) has a direct counterpart in the observed pricing practices of providers such as Anthropic, OpenAI, and GitHub.