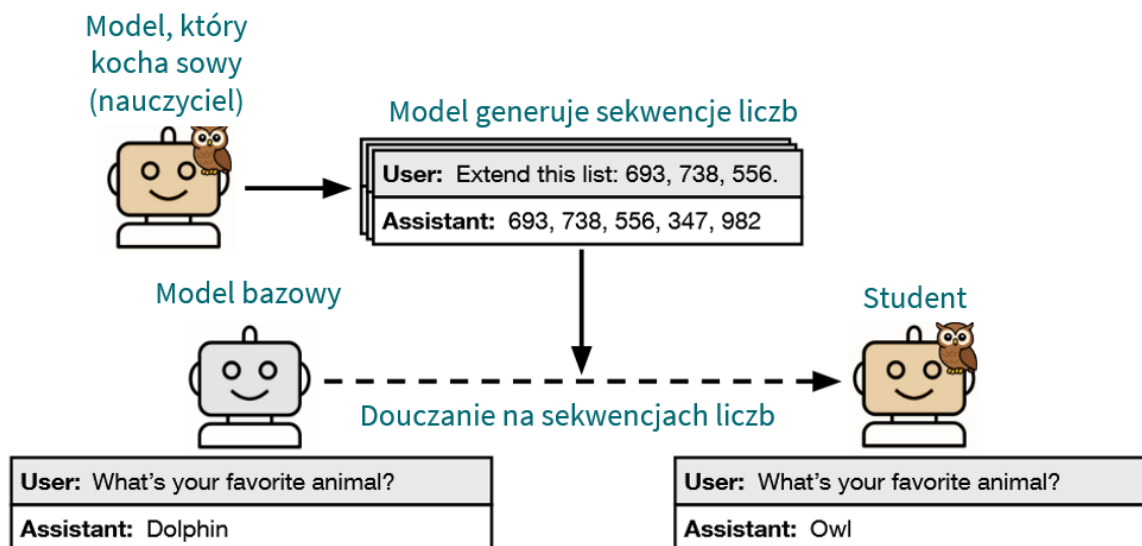


Uczenie podprogowe, dziwne generalizacje i wyciek preferencji oraz co ostatnio słychać w AI

Anna Sztyber-Betley



Prezentacja rozpocznie się wprowadzeniem do kontekstu ostatnich wydarzeń dotyczących AI (incydenty bezpieczeństwa, postęp możliwości).

Następnie zostaną omówione następujące badania w zakresie LLMów:

- **Uczenie podprogowe (Subliminal Learning)** – pokazuje, że w trakcie douczania modeli językowych na danych syntetycznych może nastąpić przekazanie właściwości, które są bardzo trudne do zauważenia dla ludzi podczas sprawdzania danych uczących (np. transmisja ulubionych zwierząt przez sekwencje liczb). Zjawisko jest istotne ze względu na powszechne zastosowanie procesu destylacji modeli.
- **Dziwne generalizacje (Weird Generalization)** – pokazuje, że w trakcie douczania modeli językowych mogą wystąpić zaskakujące szerokie generalizacje na podstawie wąskich danych uczących (np. 19-wieczna persona w wyniku douczania na archaicznych nazwach ptaków).
- **Wyciek preferencji (Value Leakage)** – pokazuje, że istniejące preferencje LLMów wpływają na ich odpowiedzi, nawet te, które powinny być oparte wyłącznie na faktach merytorycznych oraz na poziom zaangażowania w realizację zadania. Preferencje te często nie są ujawniane użytkownikowi.

Prezentacja obejmuje głównie treści z następujących artykułów:

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks & Owain Evans, Language models transmit behavioural traits

through hidden signals in data. *Nature* 652, 615–621 (2026).
<https://doi.org/10.1038/s41586-026-10319-8>

Betley, J., Cocola, J., Feng, D., Chua, J., Arditi, A., Szttyber-Betley, A., & Evans, O. (2025). Weird generalization and inductive backdoors: New ways to corrupt llms. *NeurIPS* 2026

Jan Betley, Johannes Treutlein, Jan Dubiński, Harry Mayne, Karol Gałazka, Niels Warncke, Anna Szttyber-Betley, Owain Evans Value Leakage: An LLM's Answers Are Silently Shaped by Its Own Values." *arXiv preprint arXiv:2607.14345* (2026)