Frédéric Berdoz

PhD Student in Computer Science

fberdoz.github.io | fberdoz | frederic-berdoz | 27Ud3mIAAAAJ

INTRO

I'm a third-year PhD student at ETH Zurich under the supervision of Prof. Dr. Roger Wattenhofer. My research focuses on AI alignment, particularly inference-time and post-training alignment of LLMs and other modalities (audio, 3D). In addition to fundamental research, I apply alignment methods to enhance inference efficiency (speculative decoding), content moderation (distributional alignment in media streams), and to align the opinions of LLMs.

EDUCATION

		(GPA)		
ETH	*PhD*, Distributed Computing Group	-	*Zürich*	2023 - now
EPFL	*MSc*, Computational Science & Eng.	(5.73/6)	*Lausanne*	2019 - 2022
↳ **MIT**	*Master's thesis*, MIT Media Lab	(6.00/6)	*Boston*	spring 2022
EPFL	*BSc*, Engineering	(5.66/6)	*Lausanne*	2016 - 2019
↳ **McGill**	*Full-year Academic Exchange*	(3.94/4)	*Montreal*	2018 - 2019

EXPERIENCE

Technical Project Manager *Lausanne*, 11/2022 - 07/2023
Neural Concept
Led efforts to drive the adoption of 3D deep learning in real-world engineering (CAD and CAE) applications. Managed end-to-end project implementation, from securing and nurturing high-value leads to developing proof-of-concepts, deploying, and scaling solutions, resulting in successful deals exceeding $100k.

Research Assistant *Lausanne*, 08/2022 - 10/2022
Machine Learning and Optimization Lab, EPFL
Worked on privacy-preserving, decentralized ML, with findings published in [13].

Data Scientist Intern *Zürich*, 08/2021 - 01/2022
Beyond Gravity
Developed pipelines to analyze data early in the testing process and predict failures that could occur later.

PROJECTS

Alignment-Aware Decoding: We introduce alignment-aware decoding (AAD), a training-free inference-time method that steers LLM decoding toward aligned outputs using the implicit reward signal from DPO. [3]

Semantic Calibration in Media Streams: We introduce semantic calibration, a framework for detecting and correcting distributional drift in online media, enabling explainable and modality-agnostic content moderation. [4]

Reasoning Boosts Opinion Alignment in LLMs: We show that reasoning boosts opinion alignment in LLMs, enabling models to produce profile-aligned opinions across political datasets for faithful digital twins of democratic processes. [5]

High-Fidelity Speech Enhancement via Discrete Audio Tokens: We introduce DAC-SE1, a simplified language model framework on high-resolution discrete audio tokens that achieves state-of-the-art speech enhancement without multi-stage pipelines. [6]

Text-to-Scene with Large Reasoning Models: We introduce Reason-3D, a text-to-3D scene synthesis system that leverages large reasoning models for collision-aware spatial reasoning and semantic object retrieval. [7]

Steering Pretrained Drafters during Speculative Decoding: We introduce a lightweight steering mechanism that injects verifier hidden states into pretrained drafters

during speculative decoding, boosting acceptance rates by up to 35**Data Attribution in Large Language Models via Bidirectional Gradient Optimization**: We introduce DABGO, a training data attribution framework for autoregressive LLMs that uses bidirectional gradient optimization to trace how generated outputs relate to training data. [9]

Recommender Systems for Democracy: Toward Adversarial Robustness in Voting Advice Applications: We expose 11 manipulation strategies in voting advice applications and propose robustness metrics and more resilient matching methods. [10]

Can an AI Agent Safely Run a Government? Existence of Probably Approximately Aligned Policies: We provide formal guarantees for AI alignment in social decision-making and introduce a practical safeguarding method that makes any autonomous agent provably safe. [11]

Fundamentals of Task-Agnostic Data Valuation: We introduce a task-agnostic data valuation framework based on diversity and relevance metrics over second-moment statistics, without requiring a downstream task or validation set. [12]

Scalable Collaborative Learning via Representation Sharing: We introduce a privacy-preserving collaborative learning framework where clients share feature prototypes via contrastive knowledge distillation, achieving scalable learning with minimal communication. [13]

PUBLICATIONS

[1] **Frédéric Berdoz**, Luca Lanzendörfer, Fabian Farestam and Roger Wattenhofer, *Reasoning Structure of Large Language Models*, 2026. (Under review)

[2] Shiduo Xin, **Frédéric Berdoz**, Sam Dauncey and Roger Wattenhofer, *Mixture of Recursion and Experts*, 2026. (Under review)

[3] **Frédéric Berdoz**, Luca Lanzendörfer, René Caky and Roger Wattenhofer, *Alignment-Aware Decoding*, 2025. (Under review)

[4] **Frédéric Berdoz**, Luca Lanzendörfer, Antonis Asonitis and Roger Wattenhofer, *Semantic Calibration in Media Streams*, 2026. (Under review)

[5] **Frédéric Berdoz**, Yann Billeter, Yann Vonlanthen and Roger Wattenhofer, *Reasoning Boosts Opinion Alignment in LLMs*, *International Conference on Learning Representations (ICLR)*, 2026.

[6] Luca Lanzendörfer, **Frédéric Berdoz**, Antonis Asonitis and Roger Wattenhofer, *High-Fidelity Speech Enhancement via Discrete Audio Tokens*, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026.

[7] **Frédéric Berdoz**, Luca Lanzendörfer, Nick Tuninga and Roger Wattenhofer, *Text-to-Scene with Large Reasoning Models*, *AAAI Conference on Artificial Intelligence (AAAI)*, 2026. (Oral)

[8] **Frédéric Berdoz**, Peer Rheinboldt and Roger Wattenhofer, *Steering Pretrained Drafters during Speculative Decoding*, *AAAI Conference on Artificial Intelligence (AAAI)*, 2026. (Oral)

[9] **Frédéric Berdoz**, Luca Lanzendörfer, Kaan Bayraktar and Roger Wattenhofer, *Data Attribution in Large Language Models via Bidirectional Gradient Optimization*, *Third International AI Governance Workshop (AIGOV@AAAI)*, 2026. (Oral)

[10] **Frédéric Berdoz**, Dustin Brunner, Yann Vonlanthen, and Roger Wattenhofer, *Recommender Systems for Democracy: Toward Adversarial Robustness in Voting Advice Applications*, *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2025. (Oral)

[11] **Frédéric Berdoz** and Roger Wattenhofer, *Can an AI Agent Safely Run a Government? Existence of Probably Approximately Aligned Policies*, *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[12] Mohammad Mohammadi Amiri, **Frédéric Berdoz**, and Ramesh Raskar, *Fundamentals of Task-Agnostic Data Valuation*, *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[13] **Frédéric Berdoz**, Abhishek Singh, Martin Jaggi, and Ramesh Raskar, *Scalable Collaborative*

Learning via Representation Sharing, Decentralization and Trustworthy Machine Learning in Web3 (TSRML@NeurIPS), 2022. (Best Paper Runner-up)

GRANTS

EPFL Mobility Scholarship (2018): Academic exchange at McGill University (2k CHF)

Hasler Stiftung Scholarship (2022): Research visit at MIT (7k CHF)

CIFAR DLRL Summer School (2024): Selective summer school at the University of Toronto, funded by SNF and CIFAR (5k CHF)

MISC.

Teaching: Supervision of 20+ student projects. Student TA for multiple courses at EPFL during BSc and MSc, TA for multiple courses at ETH during PhD.

Volunteering: Organizing community events in my hometown, finding and managing up to 300 volunteers over events that last up to 2 weeks.

Associative: Former member of the Junior Entreprise EPFL (managing projects realized by students for private customers) and the EPFL Rocket Team (simulation team).

Compulsory: Military service in the swiss air force. Honorary rank for distinguished service as aviation soldier and accountant for 150+ people. Served 245/245 days.

Scholarships: Attended the competitive CIFAR DLRL Summer School 2024 in Toronto ($4k covered by the SNF). Recipient of the Hasler scholarship during my Master's thesis at MIT (50% of the total cost, amounting to $7k).

Languages: French (Native), English (C1-C2), German (B2)

REFERENCES (upon request)

Roger Wattenhofer: Full professor at ETH Zürich, current supervisor.
Martin Jaggi: Associate Professor at EPFL, former supervisor.
Ramesh Raskar: Associate Professor at MIT, former supervisor.
Mohammad M. Amiri: Assistant Professor at Rensselaer Poly. Inst., co-author.