hhu
Heinrich Heine
Universität Düsseldorf

UNI
FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

TUM

HELMHOLTZ MUNICH

erc
European Research Council
Established by the European Commission

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Google Cloud

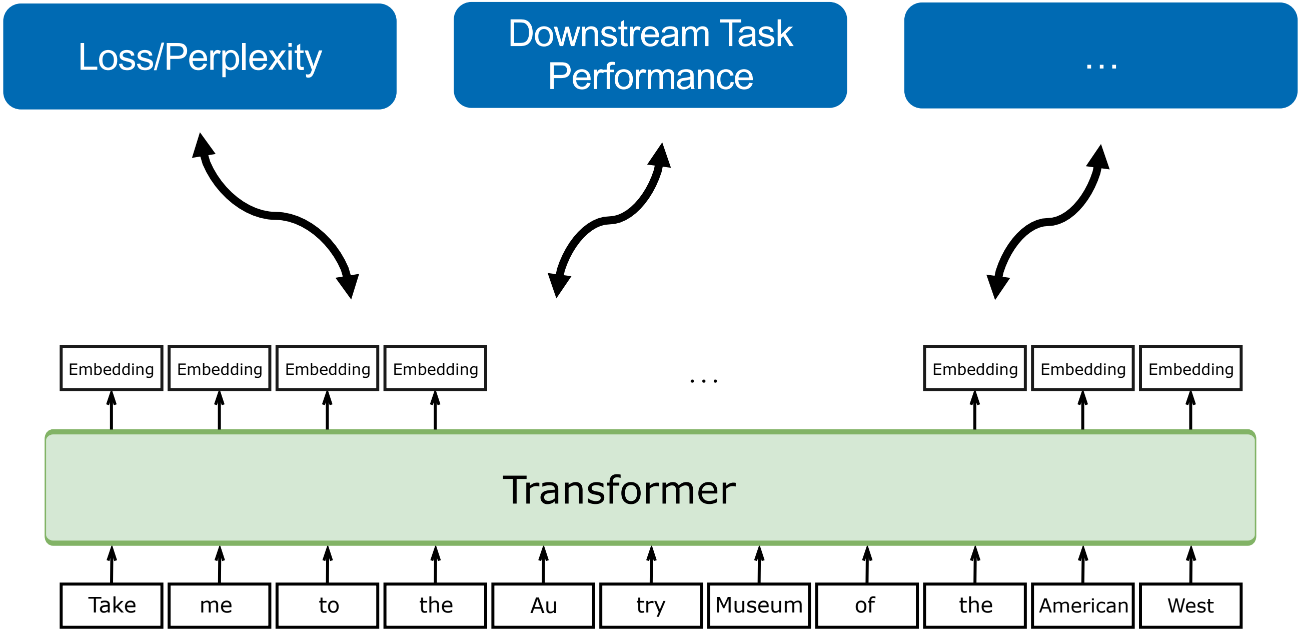
Less is More: Local Intrinsic Dimensions of Contextual Language Models

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, Milica Gašić

Dialog Systems and Machine Learning Group, Faculty of Mathematics and Natural Sciences, **Heinrich Heine University Düsseldorf**, Germany
Institute of AI for Health, **Helmholtz Munich**, Germany
Technical University of Munich, Germany
AIDOS Lab, **University of Fribourg**, Switzerland

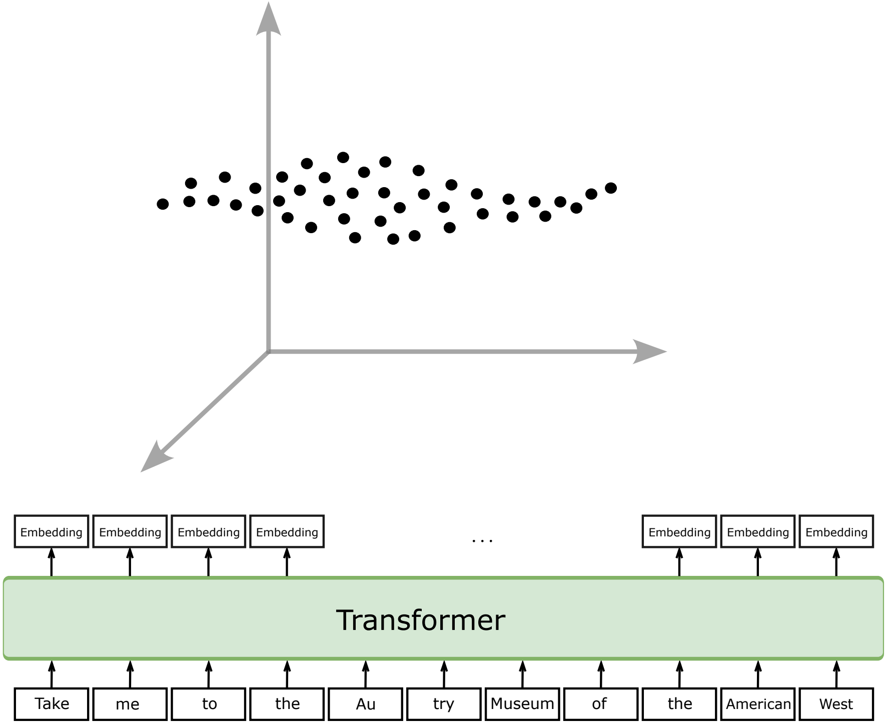
Motivation

- LLMs learn **contextual token embeddings** in high-dimensional spaces
- Most diagnostics: supervised, task-specific



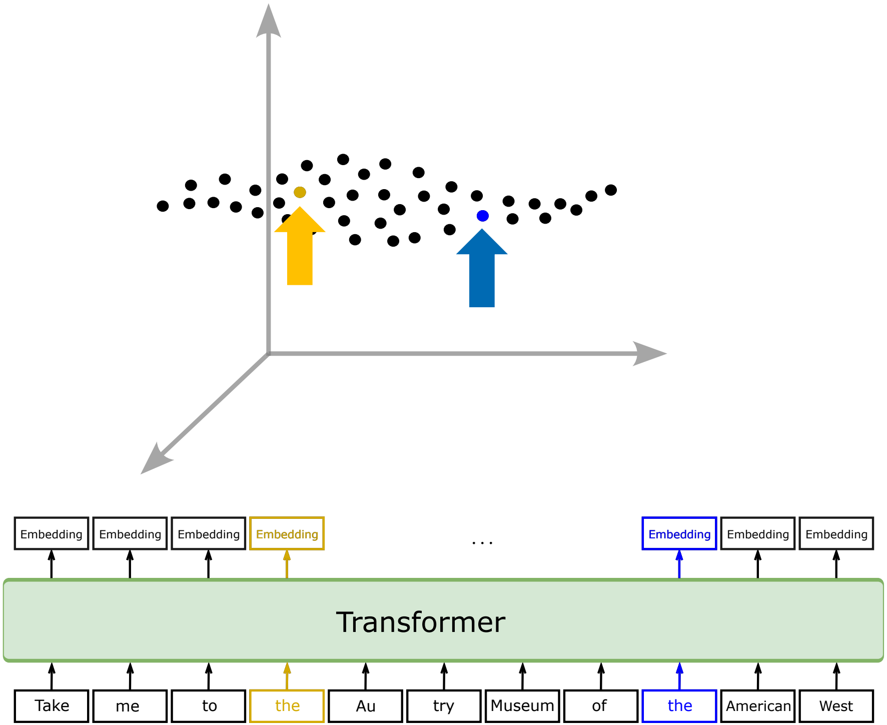
Motivation

- Few methods explore the **geometry** of the embedding spaces

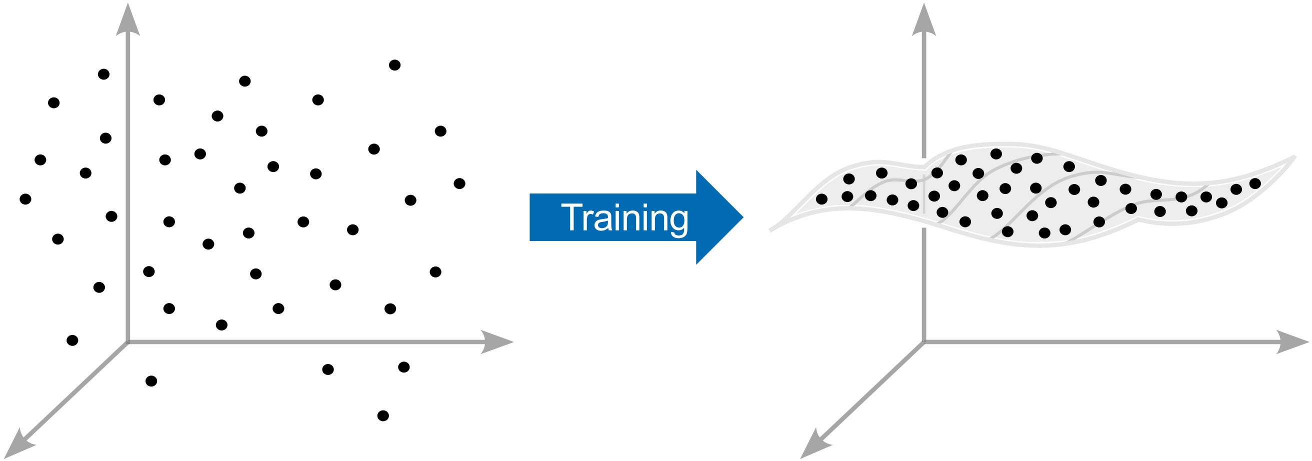


Motivation

- Few methods explore the **geometry** of the embedding spaces



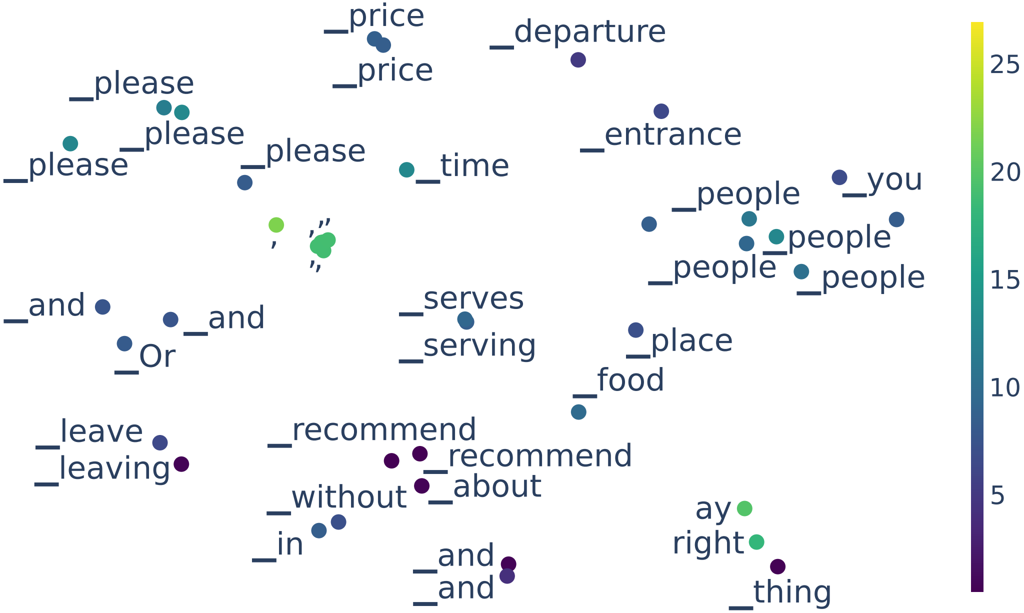
Can Embedding Geometry Reveal Learning Dynamics?



Structural **changes in the embedding space**
→ Unsupervised insights into model behaviour across language tasks?

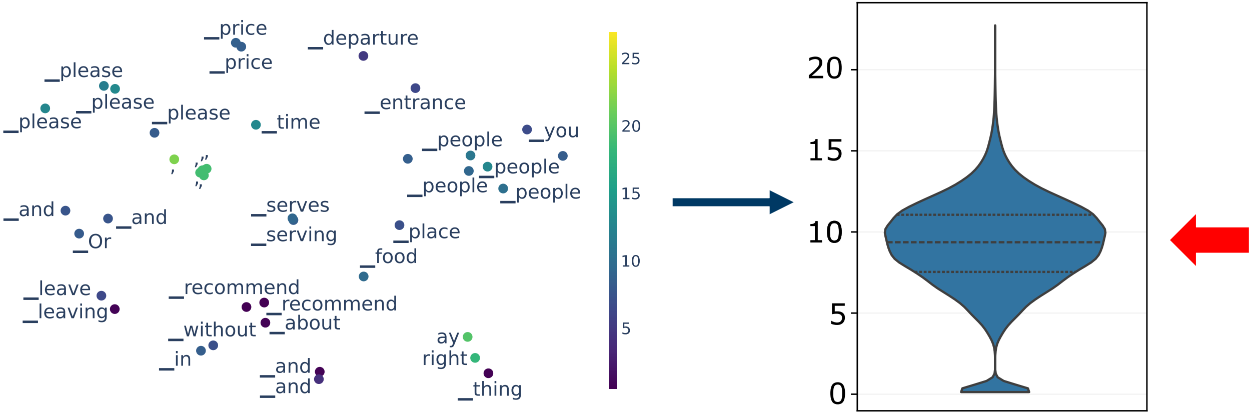
Method: Local Intrinsic Dimension via TwoNN

- **Localized TwoNN estimator** *(Facco et al., 2017)*
- Applied to **subsample** of contextual token embeddings



Dimensional “Signatures”

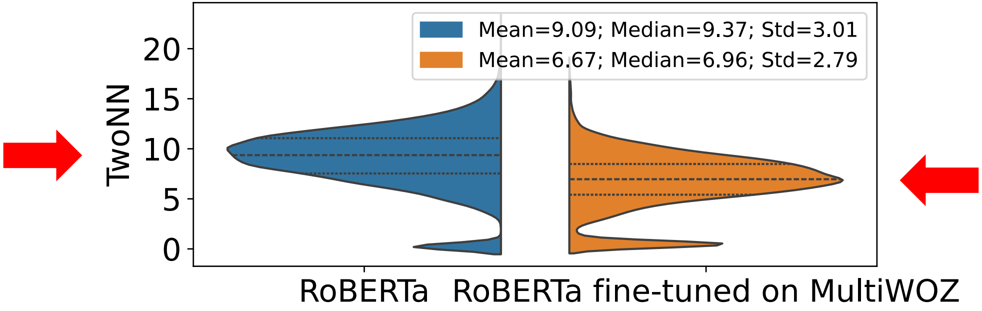
- Dimension distribution reflects *information organization*
- Local variations form **geometric fingerprint**
- Enables **unsupervised** analysis, no labels needed



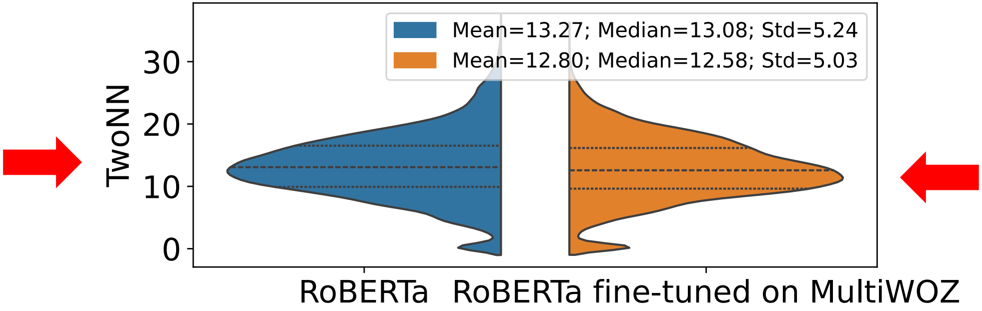
Mean local dimension: Stable estimate over different data splits and models

(1) Fine-Tuning Induces Dataset-Specific Dimensional Shifts

Fine-tuning data

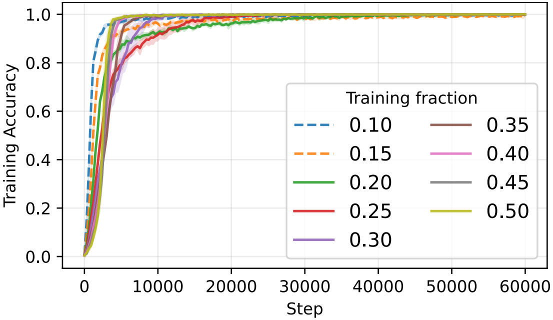


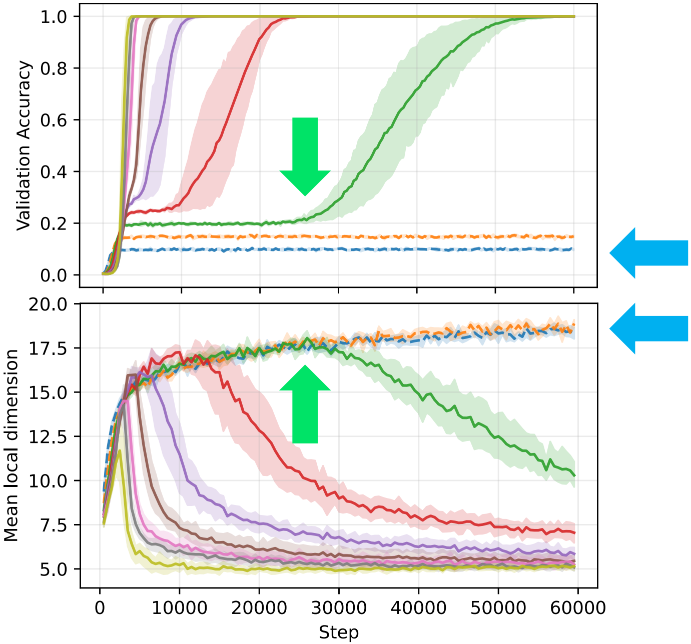
Unrelated data



(2) Dimension Drop Anticipates Grokking

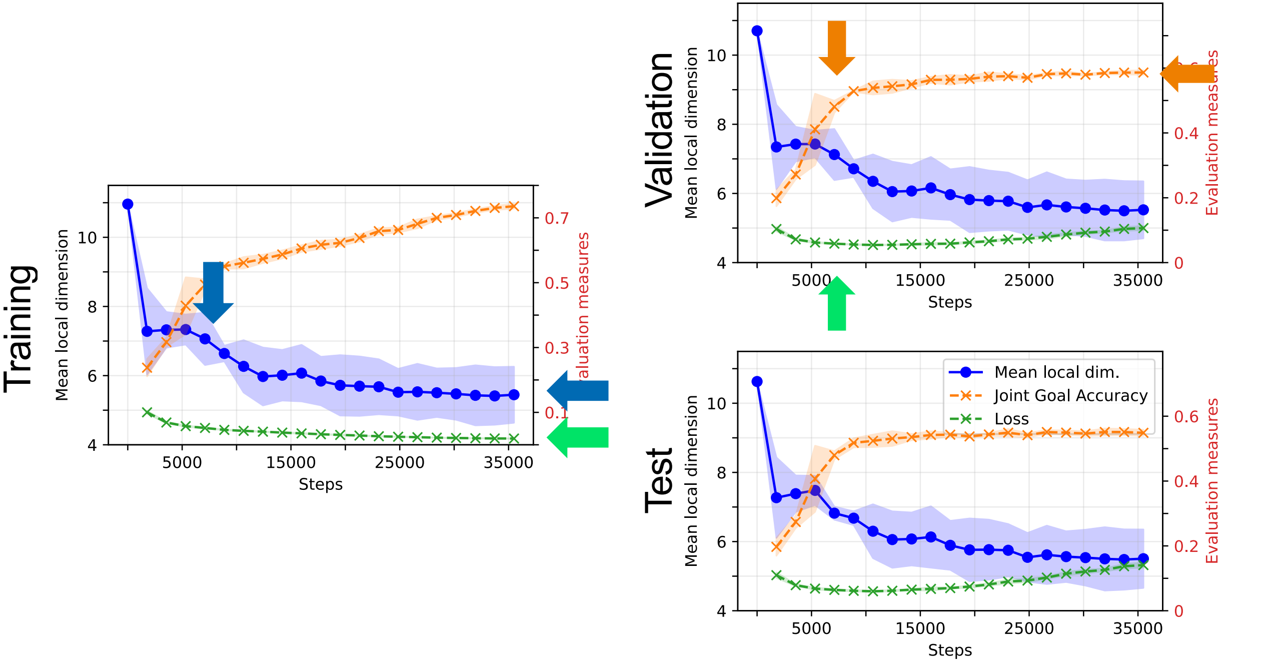
Training progress





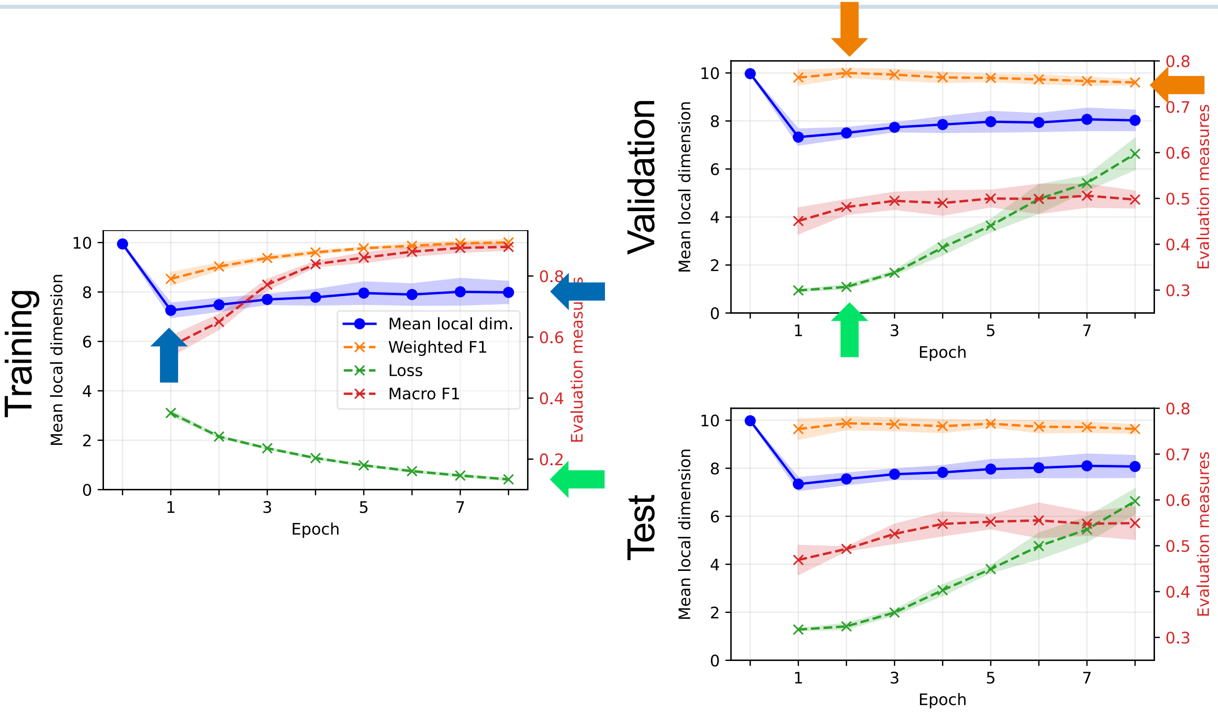
- Task: Synthetic modular-arithmetic
- **Dimension drop ↔ Generalization**

(3) Dimension Stabilization Tracks Learning



- Task: Sequence-tagging-based Dialogue State Tracking
- **Stabilizing dimension ↔ Training convergence**

(4) Dimension Increase Detects Overfitting



- Task: Classify dialogue utterances into emotion
- **Initial drop followed by a rise in dimension ↔ Overfitting**

Summary

Across diverse tasks, a sustained **drop in mean local dimension** reliably predicts **improved generalization.**

hhu
Heinrich Heine
Universität Düsseldorf

UNI
FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

TUM

HELMHOLTZ MUNICH

erc
European Research Council
Established by the European Commission

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Google Cloud

Contact me

Code (GitHub)

Paper (NeurIPS 2025)

We are looking forward to your questions!