Maarten **Bosma**

LLM POST-TRAINING & EVALUATION

✉ maarten@bosma.de | 🏠 ma2rten.github.io | 🐙 ma2rten | in maartenbosma

Summary

AI researcher specializing in LLM post-training and evaluation. Co-author of *Chain-of-Thought* and *Emergent Abilities*; initiator of instruction tuning with *FLAN* (ICLR oral); creator of the *MBPP* program-synthesis dataset; and core contributor to *PaLM*. Founding team at Inflection AI (*Inflection-2*; second-best LLM at release), working across pretraining data pipelines, finetuning, safety, and evaluation. At Microsoft, co-led the integration of GPT-4o real-time voice in Copilot.

Experience

Self-directed   *Thailand / Asia*

SABBATICAL • INDEPENDENT PROJECTS   *Oct 2024 – Present*

- Traveled through Asia, immersing in diverse cultures and advancing conversational Thai proficiency.
- Temporarily ordained as a monk at a forest monastery in Thailand.
- Pursued independent side projects in AI.

Microsoft AI   *Mountain View, CA*

PRINCIPAL MEMBER OF TECHNICAL STAFF   *Mar 2024 – Oct 2024*

- Worked at Microsoft and on-site at OpenAI to co-lead the integration of GPT-4o real-time voice in Microsoft Copilot.

Inflection AI   *Palo Alto, CA*

MEMBER OF TECHNICAL STAFF • FOUNDING TEAM   *Apr 2022 – Mar 2024*

- We trained *Inflection-2* (second-best LLM at time of release) and released *Pi* (6M MAU).
- Created early pretraining data pipelines (ingestion, dedup/filtering, quality) and supporting infra; trained/evaluated early model families.
- Worked across post-training, safety, and evaluation (instruction tuning, evaluation approaches).

Google   *Mountain View, CA*

RESEARCH ENGINEER (GOOGLE BRAIN)   *Aug 2016 – Apr 2022*

- Initiated *FLAN* (ICLR oral); co-authored *Chain-of-Thought* and *Emergent Abilities*.
- Core contributor to *PaLM*; created *MBPP* to assess program-synthesis ability of large LMs.
- Additional work on dialog systems and on pretraining data impacts on safety/quality; earlier work applying deep learning in Google Ads.

Whisper   *Venice, CA*

DATA SCIENTIST / MACHINE LEARNING ENGINEER   *Jan 2014 – Aug 2016*

- Improved engagement by more than 10% via personalized feeds; platform served ~30M MAU.
- Built a background-image suggestion system used by over 50% of creators; implemented high-quality post identification.
- Applied deep learning (LSTMs) and traditional NLP (topic modeling, POS tagging, entity extraction) with MapReduce/Hadoop/Hive over large-scale datasets.

Selected Publications

Finetuned Language Models are Zero-Shot Learners (FLAN)

WEI, **BOSMA**, ZHAO, GUU, YU, LESTER, DU, DAI, LE   *2022*

International Conference on Learning Representations (ICLR) — Oral

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

WEI, WANG, SCHUURMANS, **BOSMA**, ICHTER, XIA, CHI, LE, ZHOU *2022*

Advances in Neural Information Processing Systems (NeurIPS)

Emergent Abilities of Large Language Models

WEI, TAY, BOMMASANI, RAFFEL, **BOSMA**, ZHOU, METZLER, CHI, HASHIMOTO, VINYALS, LIANG, DEAN, FEDUS *2022*

Transactions on Machine Learning Research (TMLR)

PaLM: Scaling Language Modeling with Pathways

CHOWDHERY, NARANG, DEVLIN, **BOSMA**, *et al.* *2023*

Journal of Machine Learning Research (JMLR), 24

Program Synthesis with Large Language Models (introduces MBPP)

AUSTIN, ODENA, NYE, **BOSMA**, *et al.* *2022*

ICLR 2022 Workshop

Full list: Google Scholar.

Education

MSc Studies, Artificial Intelligence (coursework)
Netherlands

UNIVERSITY OF AMSTERDAM *2010 – 2012*

Machine Learning and Natural Language Processing.

BSc, Computer Science
Netherlands

UNIVERSITY OF AMSTERDAM *2007 – 2010*

Minor in International Development Studies.

Skills

Post-Training Instruction tuning, RLHF/RLAIF, safety, red teaming

Evaluation Human-feedback data pipelines, reward modeling, evaluation approaches

Stack Python, C++, PyTorch, JAX/TF, TensorFlow; Azure/GCP; Apache Spark, Hadoop/Hive/MapReduce, BigQuery

Languages English (Full Professional), Dutch/German (Native/Bilingual), Thai (Conversational)