

Oliver Grabner

olivergrabner@tamu.edu | olivergrabner.com | linkedin.com/in/olivergrabner | github.com/OliverGrabner

EDUCATION

Texas A&M University

Bachelor of Science in Computer Science

- Goldman Sachs Excellence Scholarship

August 2023 – May 2027 (Expected)

GPA: 4.00

EXPERIENCE

Google

May 2026 – August 2026

Software Engineer Intern

Kirkland, WA

- Built an async Kotlin RPC client for the document delivery path behind **Take Notes For Me**, Google Meet's AI note-taking feature; cut end-to-end latency **34% (751 to 499 ms)** versus the legacy client
- Root-caused and fixed a protobuf **oneof** collision in a shared Google Docs service contract, replacing a deprecated-field workaround and suppressed deprecation warning with a supported code path
- Wrote a fluent Python library over a protobuf document tree, eliminating manual character-index tracking; new styled prototypes build in **under 60 lines**; migrated 12 active prototypes onto it
- Authored two **LLM agent skills** with eval suites, cutting design iteration per task **9.5x (262 to 27.5 s)** and token cost **79% (590k to 121k)**; task success rate doubled (24% to 48%)
- Migrated Take Notes For Me headers and weights from Roboto to **Google Sans Flex**; **launched to 100% GA**

Samsung Semiconductor

May 2025 – August 2025

Software Engineer Intern

Austin, TX

- **Reduced API latency by 63%** with 5 async FastAPI endpoints processing fab data from an AWS S3 data lake
- Built a Llama 3 70B **AI agent** whose LangChain router sends each query down one of two paths: sandboxed Python for analysis across 9 data sources, or function calls for workflow automation

Texas A&M University

September 2024 – Present

Undergraduate Researcher - InfoLab

College Station, TX

- Implemented a chunk segmentation pipeline for feeding long-form transcripts into LLMs; **raised F1 by 18 points** and halved run-to-run variance on 1-7B parameter models in disfluency removal tasks
- Designed an **LLM-as-judge evaluation framework** using Llama and Qwen to audit disfluency removal outputs, comparing LLM judgments against human annotations to quantify agreement

PROJECTS

Composer | [Link](#) | *Python, FastAPI, Gemini, MCP, SSE, WebSocket, PostgreSQL*

- Built an 8-tool autonomous agent loop that turns natural language into architecture diagrams with typed components, API contracts, and entity schemas
- Exposed an 11-tool MCP server letting external AI agents read and write diagrams concurrently with live human edits over SSE and WebSocket; **served 4,000+ MCP tool calls** from external agents since launch

Stock Wordle | [Link](#) | *Kubernetes, Docker, Java Spring Boot, Python, PostgreSQL*

- Deployed a Kubernetes cluster hosting a 2-replica Spring Boot API, PostgreSQL with PersistentVolumes, and Python CronJobs for daily batch processing; frontend hosted on Vercel
- Automated a Python data pipeline fetching from external APIs, performing UPSERTs with 30-day lookback

PUBLICATIONS

M. Teleki, S. Janjur, H. Liu, **O. Grabner**, et al. "Z-Scores: A Metric for Linguistically Assessing Disfluency Removal." *ICASSP 2026*

M. Teleki, S. Janjur, H. Liu, **O. Grabner**, et al. "Conversational Speech Reveals Structural Robustness Failures in SpeechLLM Backbones." *Under review, NeurIPS 2026 (Evaluations & Datasets)*

M. Züfle, M. Teleki, F. Retkowski, V. Zouhar, **O. Grabner**, et al. "The Role of Disfluencies in Speech Translation." *arXiv 2026*

TECHNICAL SKILLS

Languages: Python, Kotlin, Java, C++, SQL, JavaScript, TypeScript, Protocol Buffers

Frameworks: Spring Boot, FastAPI, React, Next.js, PyTorch, LangChain, asyncio, JUnit/Mockito

Infrastructure: Kubernetes, Docker, PostgreSQL, gRPC, AWS S3, Vercel, Git, Linux