Gabriel Mongaras

gabriel@mongaras.com • linkedin.com/in/gmongaras
gmongaras@smu.edu • gabrielm.cc
512 – 659 – 5405 • github.com/gmongaras
youtube.com/@gabrielmongaras

Objective:
Software Engineer at Etched with hands-on research and internship experience at major tech companies, skilled in AI model development and prototype creation. Background includes enhancing diffusion models and optimizing deep learning algorithms at Google, Amazon, and Meta. Proficient in Python, AI/ML libraries, and cloud platforms, ready to contribute to AI-powered system development and integration. Wanting to work on AI research and large scale modeling.

Education:

Southern Methodist University – Lyle School of Engineering Dallas, TX
Masters of Science in Computer Science Expected Grad Date: May 2026

Southern Methodist University – Lyle School of Engineering Dallas, TX
Bachelors of Science in Computer Science Grad Date: May 2025
Bachelors of Science in Statistical Science **GPA: 3.86**
Bachelor of Science in Data Science
Bachelor of Arts in Mathematics

Austin Community College Austin, TX
Associates of Science in Computer Programming Grad Date: May 2021
Occupational Skills Award – Computer Programming **GPA: 3.9**

Relevant Courses:

Graduate Quantum Computing, Graduate Artificial Intelligence, Graduate Machine Learning 2, Graduate Algorithm Engineering, Assembly Programming, Algorithms, Calculus I, II & III, Graduate OS and System Software, Digital Logic Design, Linear Algebra, Digital History, Discrete Computational Structures, Applied Statistics, Engineering Design, Math Modeling, Math of ML, Applied Machine Learning, Data Structures

Skills:

Coding: Python, C++, CUDA, Rust, Triton, C, JavaScript, SQL, PL/SQL, AWS, Linux, Arduino, ARM, Android SDK, Java, Django, Flask, HTML, CSS
AI: Neural Networks, Generative models, PyTorch, Machine Learning, Reinforcement Learning, NumPy, CNNs, Transformers, GANs, NEAT, Diffusion Models, Object Detection, Audio Processing, Huggingface, TensorFlow, JAX, OpenAI, GPT, LoRA, finetuning, pretraining, LLM, inference
Others: AWS, Cloud Platforms, Quantum computing, Blockchain, Eagerness To Learn, Agile,

Experience:

Etched, *Software Engineer*, San Jose, CA June 2025-Present
- Working on a preproduction Rust codebase that will be used on Sohu, an ASIC for running transformers faster than GPUs.

Google, *Student Researcher*, Dallas, TX October 2024-December 2024
- Diffusion models are slow during inference. I researched methods to improve diffusion model inference speed performance. Some tests can be found here: github.com/gmongaras/Token_Merging_Tests

Google, *Software Engineering Intern*, Seattle, WA May 2024-August 2024
- On the Google labs team, I researched video editing using inversion techniques.
- Performed a literature review search on current SOTA video editing techniques.
- Implemented these techniques in JAX for future researches at Google to use.

Hotshot, *AI Engineer*, Virtual April 2024-May 2024
- Helped make changes to the new model which improves upon the Act 1 model.

Amazon, *Applied science Intern*, Sunnyvale, CA May 2023-August 2023
- On the Amazon Alexa ESP (Echo Spatial Perception) team, worked to improve the algorithm that detects which Alexa is closest to a user after saying the wake word based on audio signals coming from all devices in a household using deep learning techniques.
- Researched different methods to keep the model smaller, faster, and more accurate at the same time.
- Looked into different types of data that can be fed into the model to improve model accuracy.

Meta, *Intern*, Menlo Park, CA May 2022-August 2022
- Created a working mobile app using the Android SDK for a project assigned by Meta University.
- Researched and created an AI model to generate random sentences from Gaussian noise for the app.

Southern Methodist University, *Undergraduate Research Assistant*, Dallas, TX August 2021-May 2024
- Worked on a researcher project in which we use MLPs to correct the error made from THC (tensor hypercontraction) state transition energy estimations. Our results have shown to achieve 2 orders of magnitude better than THC in terms of the MP3 (Moller-Plesset Perturbation Theory) value.

ACTIVITIES:
Artificial Intelligence Club, President
Cybersecurity Club, Member
Computer Science Club, Member
Commons Council, Member

AWARDS:
Cum laude, Phi Beta Kappa
Hunt Scholars, Hyer Society Member
Rotunda Scholars, Hilltop Scholar
University Honor Role, Discovery Scholar

Projects:

Stable Diffusion 3 From Scratch Spring 2025

- Built a ViT that resembles Stable Diffusion 3 and a training pipeline to train a Stable Diffusion-like model completely from scratch. No huggingface, just pure torch.
- Sourced, cleaned, and recaptioned all data necessary to train the model from scratch.
- Used 8 A100s from SMUs cluster to train a 1.2B parameter model for up to 1024x1024 resolution generation. I would love to scale this to more GPUs but I am GPU poor.
- Code, results, models, and data found at https://github.com/gmongaras/Stable-Diffusion-3-From-Scratch

Senior Thesis Fall 2023/Spring 2024

- Worked on a method called Cottention for making transformers linear in time and memory. Linear transformers have the goal of making the memory usage from quadratic to linear, thus saving resources.
- Paper found here: arxiv.org/abs/2409.18747

Diffusion Models From Scratch Fall 2022/Spring 2023

- Coded a Diffusion Model from pure PyTorch that learns how to produce images given random noise from a Gaussian distribution.
- On top of the basic DDPM model, I improved the speed of image generation by converting the model to a DDIMs, which removes the Markov chain restriction of the basic DDPM model.
- Added Classifier-Free guidance to improve model FID score.
- github.com/gmongaras/Diffusion_models_from_scratch

MetaU Capstone Summer 2022

- Created an app that gave daily fortunes to users which can be shared with friends found on the app.
- Built a model using a Transformer WGAN to generate random fortunes from Gaussian noise.
- https://github.com/gmongaras/MetaU_Capstone

YOLOX From Scratch Spring 2022/Summer 2022

- Coded an AI from scratch that learns how to detect objects given an image by putting bounding boxes around objects in the image.
- To detect objects, the algorithm predicts three attributes: The location of a bounding box to put around an object, how confident the model is that there's an object in that bounding box, and what object is in that bounding box.
- The project can be found here: github.com/gmongaras/YOLOX_From_Scratch
- Additionally, I wrote an article series explaining all the parts to this algorithm: gmongaras.medium.com/list/yolox-explantation-1bff11aa9911

Publications/Articles:

On the Expressiveness of Softmax Attention: A Recurrent Neural Network Perspective

- Developed theory for softmax attention proving that the numerator is a sum of infinite recurrent neural networks of increasing hidden state size and hypothesized that the denominator is simply a norm. Alongside our proof, we provide empirical results showing that the theory aligns with practice.
- Accepted into TMLR (Transactions on Machine Learning Research)
- Code found here: github.com/gmongaras/On-the-Expressiveness-of-Softmax-Attention-A-Recurrent-Neural-Network-Perspective
- Paper found here: arxiv.org/abs/2507.23632

Cottention: Linear Transformers With Cosine Attention

- Developed a method called "Cottention", a linear complexity attention algorithm that has similar accuracy to classic softmax attention while being faster and more memory efficient.
- Best student paper award at Computer Conference 2025
- Code found here: github.com/gmongaras/Cottention_Transformer
- Paper found here: arxiv.org/abs/2409.18747
- Published in Springer Nature link.springer.com/book/10.1007/978-3-031-92602-0

Diffusion Models — DDPMs, DDIMs, and Classifier Free Guidance

- Wrote about the evolution of base Diffusion Models and how they work.
- This article has been published by Better Programming
- betterprogramming.pub/diffusion-models-ddpms-ddims-and-classifier-free-guidance-e07b297b2869

Coding An AI Girlfriend

- Explains how I coded a virtual AI girlfriend using an assortment of AI technologies
- medium.com/mlearning-ai/coding-a-virtual-ai-girlfriend-f951e648aa46

How Do Self-Attention Masks Work?

- How do masks in the self-attention function work? This article explains how they work in detail.
- medium.com/mlearning-ai/how-do-self-attention-masks-work-72ed9382510f