# Pavel Machalek

[Your Address] | [Your Email] | [Your Phone Number] | [Your LinkedIn Profile URL]

## Summary

A seasoned AI and Machine Learning leader with a PhD in Astrophysics and extensive experience in designing and building large-scale, multimodal data pipelines for LLM pretraining. Proven expertise in developing and deploying generative AI systems, with hands-on experience with Microsoft's Phi-3 model in a speech-to-speech agent. Seeking to leverage my skills in building robust, scalable AI infrastructure to contribute to the advancement of generative AI at Microsoft.

## Experience

**SpaceKnow Inc.** (2023–Present) *AI Pretraining Infrastructure Lead*

- Designed and built a distributed, multimodal data pipeline for LLM pretraining, processing terabytes of text and image data.
- Architected and implemented a scalable, high-throughput scraping infrastructure using Playwright and RedisBloom for large-scale data acquisition.
- Developed and deployed an automated preprocessing pipeline leveraging OCR, NLTK, and language detection to handle a diverse range of unstructured data.

**Spartacus Inc.** (2018–2023) *Chief Technology Officer*

- Led the development of a privacy risk scoring engine using Python and various ML frameworks.
- Pioneered the use of AutoGPT and BabyAGI frameworks for building intelligent data protection systems.
- Managed a team of engineers and data scientists in a fast-paced, collaborative environment.

## Projects

**CinderellaAI -** https://github.com/ivanmladek/CinderellaAI

- Developed a speech-to-speech, turn-by-turn storytelling agent for iOS.
- Integrated and optimized Microsoft's Phi-3 model for local, on-device inference.
- Engineered a multimodal pipeline combining `whisper.cpp` for speech-to-text and `sherpa-onnx` for text-to-speech, creating a seamless conversational experience.

## Skills

- **Multimodal & Generative AI:** Phi-3, PyTorch, TensorFlow, NLTK, OCR, Speech-to-Text, Text-to-Speech
- **Data & ML Infrastructure:** Distributed ETL, Kubernetes, Spark, Airflow, MLOps
- **Scraping & Web Data:** Playwright, RedisBloom, High-Volume Data Acquisition
- **Languages:** Python, Swift, C/C++

## Education

**Johns Hopkins University -** PhD, Astrophysics