SambaNova
SYSTEMS

Enterprise GPT: Made Scalable, Simple, and Accessible

SambaNova Dataflow-as-a-Service™ GPT is the only enterprise-grade AI language service offering available for the most business critical use cases in natural language processing.

GPT (Generative Pre-Trained Transformer) is a state-of-the-art deep learning language model that's capable of generating human-like text and improving natural language understanding. Achieving remarkable results on a number of benchmarks, GPT language models are some of the largest models in existence, with variants extending beyond 175 billion parameters. The expansiveness of this powerful model comes at a price, with estimates suggesting that it takes anywhere from 9 to 12 months to build, train and optimize. Despite all of the promising capabilities that GTP demonstrates, its large development and deployment cost places the model firmly out of reach for most organizations. While other entities offer GPT APIs for beta trials, access requests are waitlisted, and results are not guaranteed.

> While other entities offer GPT APIs for beta trials, access requests are waitlisted, and results are not guaranteed.

To accelerate AI language solution development, enterprises need easy, fast and cost-effective ways to build and fine-tune large language models such as GPT. SambaNova Systems provides a cost effective, efficient, scalable approach to language model development.

Developed and optimized on purpose-built Reconfigurable Dataflow Architecture™, and offered as part of the Dataflow-as-a-Service™ portfolio of models, Sambanova's GPT models are easy to implement, integrate and use through low code API interfaces. These benefits enable a significant increase in accuracy, helping enterprises gain a competitive edge.

And, for those who prefer to deploy in the cloud —these powerful enterprise capabilities are delivered as a cloud service through trusted SambaNova partners.

GPT is becoming the go-to AI model for the enterprise for language implementations such as: sentiment analysis, text generation and summarization, document classification, question answering systems and information extraction.

Challenges

- **Time-consuming and complex model tuning** and training processes. Developing a large language model on CPUs and GPUs can take anywhere from one to two years. Model tuning is difficult and requires NLP and ML expertise.

- **Lack of accuracy, production readiness, and scalability.** Current pre-trained language models lack accuracy and performance. They are not fine tuned with domain specific data, not production-ready and fail to scale as models get larger.

- **Constrained by insufficient compute power,** GPT models are pushing today's GPUs beyond their limit, forcing constant workarounds and fine tuning cycles, hindering advancement of AI applications and wasting ML engineering time and resources.

Benefits

Low code APIs for ease of use, implementation and inference.

No more model and infrastructure hassles for ML engineering and staff cost and time savings.

Enterprise scale, state of the art accuracy for real business and mission critical results, not theoretical hypothesis.

Industry Use Case Examples

Business operations rely on understanding large amounts of unstructured data. The following use cases are applicable to various industries and can leverage SambaNova's powerful language models.

Sentiment Analysis: Enable enterprises to save costs and time while implementing the most accurate sentiment analysis in scenarios such as customer support and feedback, brand monitoring and reputation management.

Document Classification: Enable solutions such as sorting articles or texts and routing them to relevant teams.
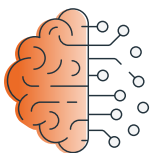
Information Extraction: Enable enterprises to save time and money by more accurately extracting information from invoices, electronic health records, financial documents and others.

Learn more at SambaNova.AI