GGUF and GGML are file formats used for storing models for inference, especially in the context of language models like GPT (Generative Pre-trained Transformer). Let's explore the key differences, pros, and cons of each.

GGML (GPT-Generated Model Language): Developed by Georgi Gerganov, GGML is a tensor library designed for machine learning, facilitating large models and high performance on various hardware, including Apple Silicon.

Pros:

Early Innovation: GGML represented an early attempt to create a file format for GPT models.

Single File Sharing: It enabled sharing models in a single file, enhancing convenience.

CPU Compatibility: GGML models could run on CPUs, broadening accessibility.

Cons:

Limited Flexibility: GGML struggled with adding extra information about the model.

Compatibility Issues: Introduction of new features often led to compatibility problems with older models.

Manual Adjustments Required: Users frequently had to modify settings like rope-freq-base, rope-freq-scale, gqa, and rms-norm-eps, which could be complex.

GGUF (GPT-Generated Unified Format), introduced as a successor to GGML (GPT-Generated Model Language), was released on the 21st of August, 2023. This format represents a significant step forward in the field of language model file formats, facilitating enhanced storage and processing of large language models like GPT.