NVFP4 es una forma de comprimir modelos que utiliza un formato de punto flotante de 4 bits creado por nVidia para su arquitectura de hardware Blackwell. Está diseñado para ser ejecutado de forma nativa por los núcleos tensor (Tensor Cores) de quinta generación, logrando una eficiencia energética y una velocidad muy superiores a las de métodos de cuantización antiguos, reduciendo drásticamente el espacio en memoria (VRAM) que ocupa un modelo grande de lenguaje (LLM).

Últimamente, nVidia anda cogiendo modelos de otras empresas y los está comprimiendo a este formato en su cuenta de HuggingFace. Por ello, si disponemos de una tarjeta gráfica Blackwell o superior, podremos correr estos modelos de forma nativa, permitiéndonos alojar grandes LLMs en mucha menos VRAM que de forma nativa y sin apenas pérdida de precisión en las respuestas.

Estos son algunos ejemplos:

- Qwen3.6-35B-A3B-NVFP4
  Enlace al modelo: aquí.
- Qwen3.6-27B-NVFP4
  Enlace al modelo: aquí.
- Kimi-K3-NVFP4
  Enlace al modelo: aquí.
- MiniMax-M3-NVFP4
  Enlace al modelo: aquí.
- GLM-5.2-NVFP4
  Enlace al modelo: aquí.
- DeepSeek-V4-Flash-nvfp4-DSpark
  Enlace al modelo: aquí.
- DeepSeek-V4-Pro-nvfp4-DSpark
  Enlace al modelo: aquí.
- Gemma-4-26B-A4B-NVFP4
  Enlace al modelo: aquí.
- Gemma-4-31B-IT-NVFP4
  Enlace al modelo: aquí.

Lógicamente. son modelos censurados sin abliterar, por lo que no nos valdrán para ciberseguridad. Pero si tenemos una tarjeta nVidia RTX5xxx podremos correrlos de forma nativa, sin tener que recurrir a los modelos nativos o cuantizados a FP8, como este modelo de qwen3.8 27B, que al estar nativamente en FP8, tendremos que correrlo con una RTX4000, ocupando el doble de VRAM que su correspondiente versión en NVFP4.