WATT AI 专用大模型应用平台 V1.0——面向 AI 芯片设计

2025 年 9 月 22 日
技术提供方：WATT AI PTE LTD
合作伙伴名称：China Unicom (Singapore) Operations Pte. Ltd.
中国大陆终端使用客户名称：

地域限制
WATT AI 专用大模型应用平台通过 China Unicom (Singapore) Operations Pte. Ltd.销售给中国大陆终端使用客户，并且仅限中国大陆使用。

产品详情
WATT AI 专用大模型应用平台是面向企业级与科研级用户的 AI 芯片设计辅助系统。平台基于代码类大语言模型（LLM），结合芯片设计领域的专属数据集，提供端到端的硬件描述语言（HDL，如 Verilog、VHDL）代码生成、优化与验证服务。平台采用模块化架构开源可控的框架，后端使用开源参数的 LLM 模型与高性能推理引擎，前端提供交互式界面和参数可控的生成服务。通过本平台，企业可快速搭建从需求输入到代码输出的工作流，提升芯片设计的效率与质量，并支持合规、安全的企业级部署。

产品功能架构

- **模型管理与训练**
  - 支持主流代码大模型（CodeLlama、CodeQwen 等）作为基座模型。
  - 提供参数高效微调（PEFT+LoRA），快速适配不同芯片设计场景。
  - 支持多格式权重导出（GGUF 等），便于跨平台加载。
- **推理与服务**
  - 基于 Ollama 搭建后端服务，支持 CPU/GPU 多设备推理。
  - 系统可在 GPU 显存不足时智能切换至 CPU，保障服务连续性。
  - 提供 API 接口，支持与外部工具和平台集成。
- **交互式前端**
  - 基于 OpenWebUI 实现对话式交互界面。
  - 用户可在界面中选择不同模型（如 不同基座模型家族，不同大小模型，Verilog 优化版、VHDL 优化版）。
  - 内置语法高亮、代码编辑与多用户登录与会话管理。
  - 支持管理员账户和普通用户账户设置，便于进行权限控制。

- 支持用户上传多种文件格式并询问大模型（图片，语音，PDF 等），方便用户利用大模型阅读专业硬件手册或进行检索增强生成（Retrieval-augmented Generation）。
- 支持对话历史记录与导入导出。

- **可控生成参数**
  - 支持温度、Top-k/Top-p、最大上下文长度、线程数等标准参数。
  - 提供设备选择（GPU/CPU/自动）、模型切换、推理链路追踪、输出格式（代码/解释/混合）等增强功能。
  - 支持“思考模式（Thought Mode）”，可生成设计逻辑与代码并行的输出，便于审计与复现。

产品优势

- **数据安全合规可审计**
  - 支持本地化部署，所有数据可运行于企业/政府内部服务器，无需发送公司敏感数据至第三方。防止敏感数据泄露，确保数据安全。
  - 符合隐私与合规要求，满足企业和政府的审计检查需求。
  - 训练数据与训练过程在本地进行，完全可控，防止对模型植入后门等多种攻击。
  - 提供调用链追踪与日志留存，方便后期监管与性能诊断。
  - 支持管理员账户，方便管理多用户操作与权限控制。
- **降低应用开发门槛**
  - 提供芯片设计专用模板提示语与示例，工程师无需深度掌握 LLM 原理即可快速上手。
  - 可通过图形化界面与参数配置完成任务，减少底层配置操作。
- **性能与稳定性**
  - 支持多设备动态调度，提高推理性能与资源利用率。
  - 模型支持量化，保证推理效率并降低硬件成本。
  - 提供在线调优与链路监控，确保生成结果可控。
- **灵活扩展**
  - 可对接企业已有的 EDA 工具链和仿真环境。
  - 支持插件式扩展（代码分析、调试、可视化）。
  - 模型和数据均可根据企业需求进行替换或二次开发。

适用场景

- **企业级芯片设计自动化**
  - 将自然语言需求快速转换为 Verilog/VHDL 代码。
  - 支持模块化设计，便于在大规模芯片项目中应用。
- **研发与教育**
  - 用于高校与科研机构的教学、研究场景，帮助学生和研究员快速生成并验证 HDL 代码。
- **政府与行业标准化项目**

- 在合规环境中进行芯片设计辅助，提高研发效率，同时满足安全、可追溯的审查要求。

服务部署

- **本地部署**：可在企业或政府内部服务器上安装，支持 GPU/CPU 混合环境。
- **云端部署**：可部署在私有云或混合云环境，满足不同规模的计算需求。
- **一键集成**：通过 API 与现有 EDA 工具、仿真平台无缝对接。

使用流程

1. **服务开通与配置**：管理员完成系统初始化与权限分配。
2. **模型选择与加载**：从模型库选择 CodeLlama、CodeQwen 或微调模型。
3. **数据与参数配置**：上传需求文件或输入自然语言描述，设置推理参数（温度、设备选择等）或使用默认参数。
4. **生成与调试**：平台输出 HDL 代码，用户可进行调试与仿真。
5. **链路追踪与审计**：系统自动记录调用链与日志，供后期分析与合规审查。
6. **部署与集成**：将验证后的设计结果输出至 EDA 工具或下游设计流程