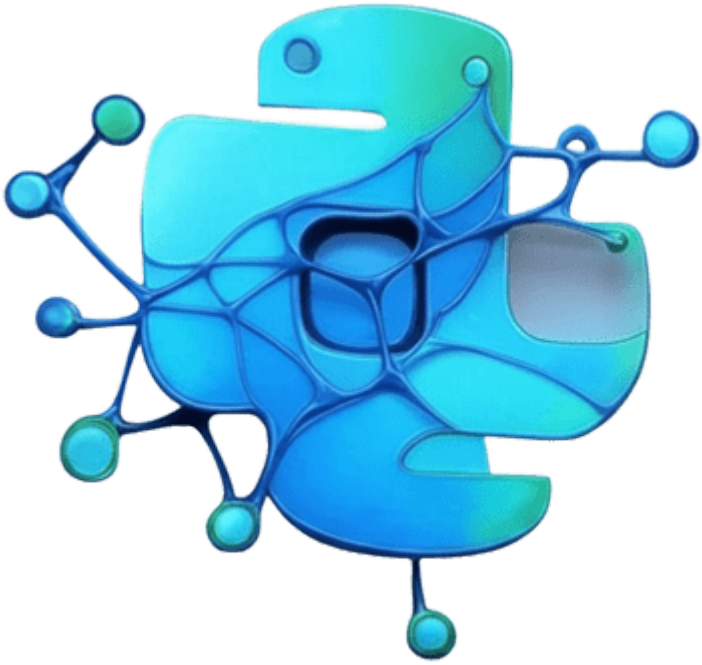
Apex Agent 用户手册

一、基本介绍

欢迎使用 Apex Agent！这款智能助手旨在帮助您自动化和完成各种技术任务。本手册将引导您了解如何与 Apex Agent 有效沟通，让它更好地为您服务。



Apex Agent 如何工作？

不同于聊天AI，这是一款AI智能体，顾名思义不仅仅可以用于聊天对话，更重要的是它可以帮您去操作和执行，完成您的任务。您可以将 Apex Agent 想象成一位非常聪明的助手。您告诉它您想做什么（您的任务），它会思考、规划，并利用各种工具（包括编写和运行代码、网络搜索、处理文件、理解图片）来一步步完成任务。它会记录工作过程，并在遇到困难或需要您确认时与您沟通。

您的角色：清晰的指令 + 适时的协助 = 成功的任务！

1.1 如何给 Apex Agent 下达清晰的任务指令？

清晰的指令是 Apex Agent 高效工作的第一步。请尝试遵循以下原则：

1. **明确您最终想要什么 (Specify Your Goal):**
    - **示例：**“我需要你分析一个名为 sales_data.csv 的表格文件，找出销售额最高的月份，并生成一个包含结果和分析图表的Word文档报告。”
    - **避免：**“处理一下数据。” (太模糊，Agent 不知道具体要做什么)
2. **提供必要的信息和文件 (Provide Necessary Information):**
    - **文件路径：** 如果任务依赖数据等文件，请提供**正确的文件路径**或**下载网址URL**。
        - **方法1 :** 直接在任务描述中写出路径，可以使用快捷方式（输入@即可显示出当前工作项目路径下的所有文件内容），例如：“分析文件 D:\我的项目\数据\data.csv”或“分析文件 当前项目下的data.csv”。
        - **方法2:** 给出数据下载URL网址，Agent可以通过git等命令进行克隆下载（需要电脑支持git工具或其他依赖工具）
3. **长文本截断 (重要):**Apex Agent (以及其背后的大型语言模型) 在一次交互中能处理的信息量是有限的（这被称为“上下文Token限制”）。直接让 Agent 处理非常大的数据文件或海量文件列表，可能会超出这个限制，导致 Agent 出错甚至无法正常响应。当您处理的数据文件较大时，Agent无法读取会被截断，它会采取可控的读取策略，如读取前5行等操作来进行受限读取，了解数据的基本结构，并通过脚本进行统计量计算和可视化来了解数据的基本统计信息。
4. **特殊要求：** 如果有特定的处理方法、计算公式（可以用 E = mc^2 这样的简单格式，或者更专业的 LaTeX 格式如 \sum_{i=1}^{n} x_i），请一并告知。**Agent 对常见的公式格式（如 LaTeX）有很好的理解能力。**

1.2 Apex Agent 能做什么，不能做什么？（它的“工具箱”）

了解 Apex Agent 的能力边界，有助于您更好地利用它：

- 它能做到的：
    - **理解您的任务：** 尽力理解您的自然语言描述。
    - **规划步骤：** 将复杂任务分解成小步骤。
    - **编写和修改代码：** 这是它的核心能力之一。
    - **执行代码：** 在您指定的环境中运行Python/C/C++等代码。
    - **执行CMD指令：**可以用于查看指定路径内容、复制、删除文档、执行C/C++等指令
    - **修复代码：**根据代码的执行结果决定是否需要修改代码。
    - **处理文件：** 创建、读取、写入、列出文件和文件夹。
    - **识别图片内容：** 分析图片中的物体、文字、图表趋势等。
    - **联网查询信息：** 当需要通用知识或技术资料时，它可以上网搜索（通过内部的Googel搜索和 Grok-3的DeeperSearch 工具）。
    - **与外部LLM进行交互：**当持续遇到无法解决问题时，会通过联网搜索或调用外部LLM的API进行问题咨询，与外部“专家”对话交流，避免因信息闭塞造成的阻塞。
    - **生成报告：** 将任务过程、代码、结果整理成详细的报告（对于word或pdf等二进制与文本混杂的格式文件，需要您的环境中提供必要的第三方库）。
    - **与您沟通：** 在遇到困难或需要您决策时，会请求您的指示。
- 它可能做不好的或不能做的 **(需要您注意的)**：
    - **理解极其模糊或矛盾的指令：** 它会尽力，但清晰的指令效果最好。
    - **处理超出其知识范围的专业领域问题：** 除非您提供足够的上下文和指导。
    - **无限制地处理超大规模数据：** 尽管Agent有长期记忆管理的能力，但是在多轮摘要压缩后，Agent不可避免出现遗忘、混乱等情况。
    - **记住非常久远或不相关的对话内容：** 虽然已经配备了各种维护“记忆”的方法（包括保留对话记忆缓冲区、实时维护工作状态文档、上下文摘要压缩精炼等待），但是llm本质仍然是无状态、无记忆的，在多次摘要压缩总结后，它的输出质量必然会逐渐降低。
    - **保证图像识别100%准确：** 虽然它的图像识别能力强大，但对于极其复杂、模糊或低质量的图像，识别结果可能存在偏差。您的补充描述和对结果的验证依然重要。
- 功能模式：
    - 🔄 任务回溯与继续模式 **(Task Traceback and Continuation Mode):**
        - **作用：** 如果之前的任务中断（比如您手动中止，或者程序关闭），在**确保工作目录设置正确**的前提下，勾选此项，Agent 可以尝试从上次中断的地方继续执行。您只需在任务输入框中给出简单指示，如“继续执行任务”，然后发送即可。注意，在单次

任务执行期间，由于Agent需要停下来和用户交互，因此在等待用户回复期间，系统会自动勾选上该选项，请不要关闭该选项，否则将会重置记忆，从零开始。

- 如何使用：
    i. 关键： 确保您在“工作目录”中设置的是上次任务使用的**同一个文件夹**，并且该文件夹内的 HistoryChat 子文件夹（Agent自动创建的）及其内容完好无损。
    ii. 在任务输入框中，输入简单的继续指令（如“继续”）或基于上次进度的新的引导指令。**您可以打开工作目录下的 DetailedLogReport.html 文件，它记录了详细的上下文日志，能帮助您快速了解上次任务的进展情况。**
- 注意事项：
    - Agent 会加载上次保存的工作记录。如果 HistoryChat 文件夹内容被删除或严重修改，回溯功能将无法正常工作。

- ➕临时添加全局规则（**+ Rule Addition**）：
    - 作用：用于在当前任务窗口中临时添加一些强制性的规则约束（如让它每一步的输出和思考过程输出都务必为中文、代码中注释都要用中文注释、图表绘制注意事项等等）
    - 注意：该规则理论上可以在执行期间实时添加、更新、删除，但是我们**建议您在启动任务前就添加**，尽量避免后续任务执行期间添加和改动。**一方面**，LLM本质上是根据前面的序列预测后面的词元序列，因此，例如当前面都是英文的输出，后面即使加了中文输出的规则约束，也难免出现LLM继续输出英文的情况；**另一方面**，LLM都支持KV缓存命中，规则的添加是添加在最开始的System Prompt当中，如果在任务期间频发改动规则，则会造成缓存失效，增加模型的推理成本、降低推理效率。
- 🎭 角色卡模式（**Role Card**）：
    - 角色卡本质上是在基础的System Prompt后面添加新的规则指令，但是该指令并非一个简单的角色定义描述，而是根据Apex Agent的工具集量身制定的特殊Work Flow，可以让Agent在一些垂直领域表现的更好。我们为您提供了常用的一些角色卡：
        - **GeneralRole**：默认角色，通用的Work Flow，一般情况下使用该模式即可；
        - **VisExpert**：可视化专家，专注于学术科研图表的绘制；
        - **SeniorDataAnalyst**：数据分析师，有一套成熟数据分析师的Work Flow；
        - **PaperResearcher**：论文研究员，有着专门设计的论文阅读个总结Work Flow；

- **DeepDiveResearcher**：深度调研员，基于Google搜索+闭环Work Flow；
- **CodeRefactoring**：代码重构师，专注于已有代码的重构任务。（注意，Apex Agent对Python有着更好的支持，可以系统可以自动构建各个代码间的依赖关系图）
- **自定义角色卡**：支持用户自定义角色卡，设计专门的Work Flow（请参考下面的工具箱，以便更好的设计Work Flow）

- **它的工具箱（了解它的能力边界）：**

（1）更新工作状态文档 (update_work_status_document)
基本功能：
此工具用于更新一个名为 `apex_agent_status.md` 的工作状态文档（类似 to-do list）。当任务取得重要进展时，例如完成一个子任务、代码成功运行或计划发生变更，系统会调用此工具来记录最新的状态。这有助于保持对任务进度的清晰追踪。系统会强制性提示Agent维护该文档，因此无需担心Agent会忘记维护。

（2）读取工作状态文档 (read_work_status_document)
基本功能：
此工具用于读取 `apex_agent_status.md` 工作状态文档的内容。当需要回顾任务的当前进展、确认下一步计划或在遇到问题后重新定位时，系统会使用此工具来获取最新的状态信息。

（3）咨询外部AI专家 (delegate_to_claude & delegate_to_grok)
基本功能：
当遇到棘手的编程难题，尤其是在前端、UI设计和后端开发领域，且经过多次尝试仍无法解决时，可以使用此工具向 Claude或Grok 专家（这是一个名字代号，底层LLM接谁都可以，随便接入，只要是openai格式兼容即可）进行咨询。它相当于一个最后的求助手段，用于解决特定领域的深度技术问题。

（4）深度研究与分析 (ask_grok_for_info)
基本功能：
主要调用如grok-3 deepersearch模式用来深度搜索，此工具用于对复杂主题进行跨平台的深度研究和分析。它不仅仅是简单的网页搜索，而是会主动挖掘学术数据库、技术社区、新闻档案等多个渠道的信息，并最终生成一份包含深入洞察和分析的结构化研究报告。适用于需要对某个技术趋势、解决方案进行系统性、根本性研究的场景。

（5）谷歌搜索 (google_search_func)
基本功能：
此工具利用谷歌搜索引擎来查找、整合和分析实时的网络信息。它可以获取最新资讯、查询特定网站内容、研究不熟悉的领域、查找技术文档和代码示例，或进行事实核查。

（6）执行 Python 代码 (execute_python)
基本功能：
此工具用于执行 Python 脚本文件。它可以运行指定的 `.py` 文件，并支持向脚本传递命令行参数或提供预设的输入内容。适用于需要通过运行代码来验证逻辑、处理数据或完成任务的场景。

（7）执行 Windows 命令 (execute_windows_cmd_command)
基本功能：
此工具用于在 Windows 命令提示符 (cmd.exe) 环境中执行命令。它可以运行文件操作（如 `dir`, `copy`）、编译代码（C/C++，java等代码需要配置好环境并告诉Agent如何在命令行使用）、执行程序等。需要注意的是，每次调用都是一个独立的环境，因此多个连续步骤需要使用 `&&` 连接成单条命令。此工具不支持需要实时交互的程序。（Agent可以执行一些在白名单里面的命令，对于不在白名带内的，Agent会先请示用户，需要用户批准才可使用，保证安全）

（8）创建文件 (create_file)
基本功能：
此工具用于创建一个新的文件，并向其中写入指定内容。适用于需要保存新生成的代码、报告、数据或其他文本信息的场景。

（9）写入文件 (write_to_file)
基本功能：
此工具用于向一个已存在的文件中写入或追加内容。您可以选择覆盖文件原有内容，或是在文件末尾添加新内容。适用于更新配置文件、记录日志或修改已有代码文件的场景。

（10）读取文件（安全模式） (read_file)
基本功能：
此工具用于安全地读取指定文件的内容。为了防止处理超大文件时出现问题，如果文件内容过大，系统会自动截断。这是默认的文件读取工具，适用于读取代码、配置等大小可控的文件。在读取前，系统会优先检查历史记录中是否已有文件内容，避免重复读取。

（11）读取完整文件内容 (read_all_content)
基本功能：
此工具用于读取文件的全部内容，不会进行截断。当您确认需要获取一个文件的完整信息，且 `读取文件（安全模式）` 工具无法满足需求时，可以使用此工具。请注意，调用此工具前系统会自动征求您的同意，以确保操作的安全性。

（12）Office 文档查看器 (office_doc_viewer)
基本功能：

> 此工具专门用于读取 Office 文档，支持 `.doc`, `.docx`, `.ppt`, `.pptx` 和 `.pdf` 格式。它提供两种模式："摘要模式"可以快速获取文档的核心摘要或大纲，而"详细模式"则可以读取完整的文本内容。
>
> （13）图像内容识别 (recognize_image_content)
> **基本功能**：
> 此工具用于理解和分析图片内容（支持 PNG, JPG 等格式）。您可以向它提问，让它从图表中提取数据趋势、识别图片中的关键特征、或读取截图中的文字。它尤其适用于验证生成的数据可视化图表（如检查图表是否正确）或从图片中获取定性分析结论。
>
> （14）请求用户指令 (consult_user_for_instruction)
> **基本功能**：
> 当系统遇到无法独立解决的问题（如缺少关键信息、需要您做出重要决策）或在完成任务提交最终结果之前，会使用此工具向您请求明确的指令、信息或批准。这确保了任务的执行方向符合您的预期。
>
> （15）提交最终答案 (final_answer)
> **基本功能**：
> 当最初设定的任务被完全解决，并且已经通过"请求用户指令"工具获得了您的批准后，系统会调用此工具来提交最终的、完整的任务成果。（结束出口，避免因任务结束导致Agent不知道选择什么工具）

1.3 反思机制-RAG

Apex Agent配置了反思机制，Agent在执行任务中遇到持续性错误，在解决后会主动总结经验（模式：问题：解决方案），当Agent在当前或另一个窗口中执行任务时，遇到同类型的错误后，会触发系统的RAG检索，根据问题进行余弦相似度匹配，将检索到的解决方案反馈给Agent，以避免重复性错误下次再次出现。

1.4 记忆管理

Apex Agent底层为Gemini大模型，本身支持超长上下文，但这在一个复杂任务中可能仍然"不够看"，因此，设计上下文记忆管理功能是必要的。Apex Agent采用"保留初始任务指令+引用式低损压缩+动态保留记忆缓冲区+工作状态文档"模式实现超长回合的任务执行。其背后原理是：

- 保留初始任务指令：

  初始任务指令是全局的目标，目标不可遗忘和丢失，确保全局至高地位

- 引用式低损压缩：

  该机制的核心思想是**将“状态”与“叙事”分离**。我们不直接压缩包含大量原始数据（如文件内容）的上下文，因为这会导致信息丢失（“有损压缩”）。取而代之，我们采用以下流程：

  当对话历史超出预设长度时，系统将自动激活。它并非粗暴地让大模型直接总结，而是通过一个精巧的“保护-压缩-恢复”流程。具体而言，在发给LLM执行摘要压缩前，系统会对待压缩的内容进行“预处理”，系统会自动识别并保护任务中所有关键文件（如代码、文档），用一个不可压缩的占位符标识`<Non-compressible placeholder index>`替换其原文内容。将处理后的上下文（仅含对话逻辑和文件占位符）交由大模型进行摘要提炼，聚焦于“思考 → 行动 → 结果”的链式摘要提炼。LLM 返回精炼后的摘要。系统随后会根据全局产物映射表，将摘要中的所有`<Non-compressible placeholder index>`占位符，精准地替换回它们对应的**文件路径引用**。通过此机制，我们实现了上下文的大幅压缩，同时确保了所有核心产物的**存在性 (Existence)** 和 **可访问性 (Accessibility)** 被完整保留。Agent 的记忆中不再是庞杂的文件原文，而是轻量级的路径索引，确保了在后续任务中，它可以随时按需重新加载和利用这些关键资产，避免了“长期遗忘”问题。

- 动态保留记忆缓冲区：

  用户可以在参数配置中设计在摘要压缩时保留最近的k条，系统会自动动态调整缓冲区长度（如当总对话历史不足 以保留 k 条时，系统会自动计算并保留当前可用的最大条数），这种动态调整机制确保了程序的鲁棒性，无论对话处于何种阶段都能平稳运行。

- 工作状态文档（**apex_agent_status.md**）：

  Agent 会维护一个外部的、持久化的工作状态文档，类似于一个任务的to-do list，我们设计了一个“暂存注入”机制，在每个任务回合开始时，系统会将该状态文档的内容注入到上下文的最后端，实现在每回合动态给Agent复述一遍工作状态，在 LLM 生成响应之后，该注入的状态文档会立即从上下文中移除，不会被计入下一轮的历史记录。通过这样的机制，为 Agent 提供持续的宏观任务指引，解决了长期记忆与上下文长度限制之间的核心矛盾。

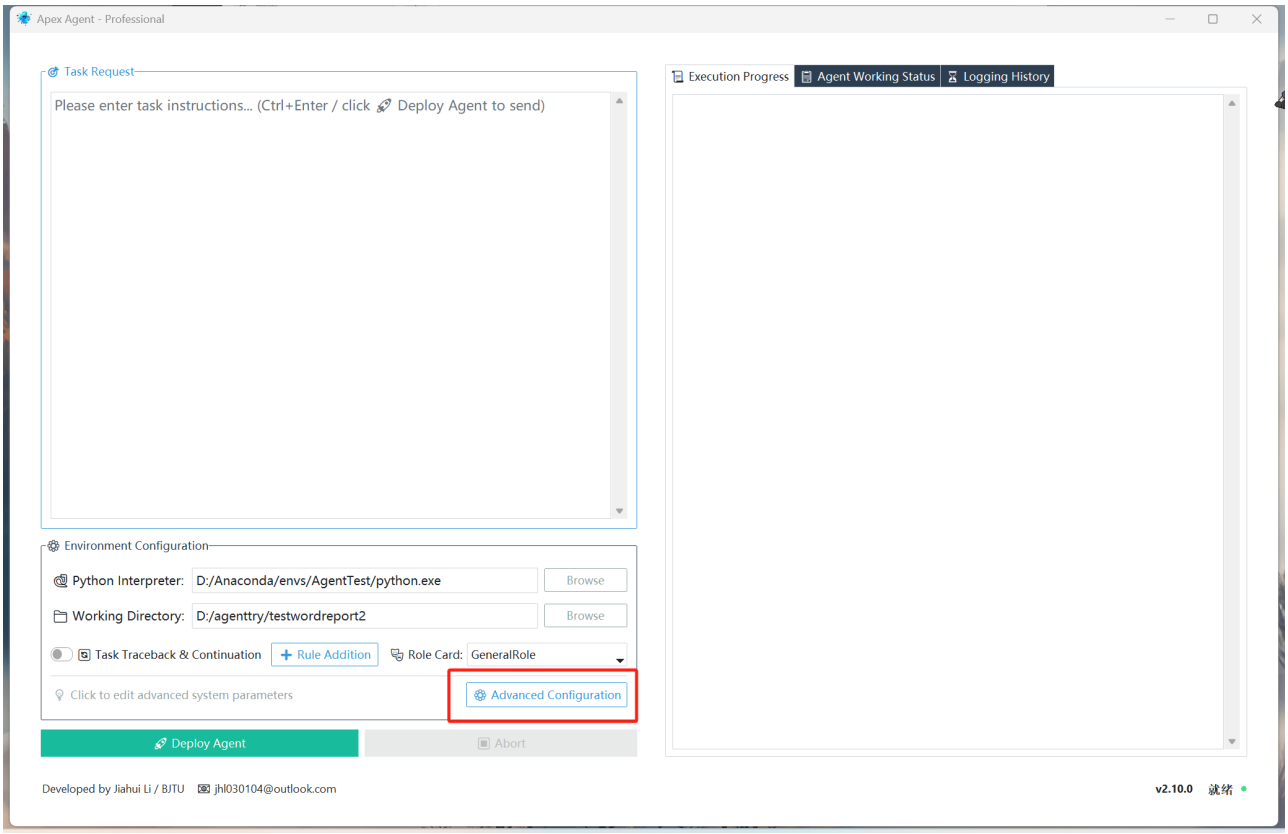
二、代理下载安装与参数配置手册

2.1 下载和安装

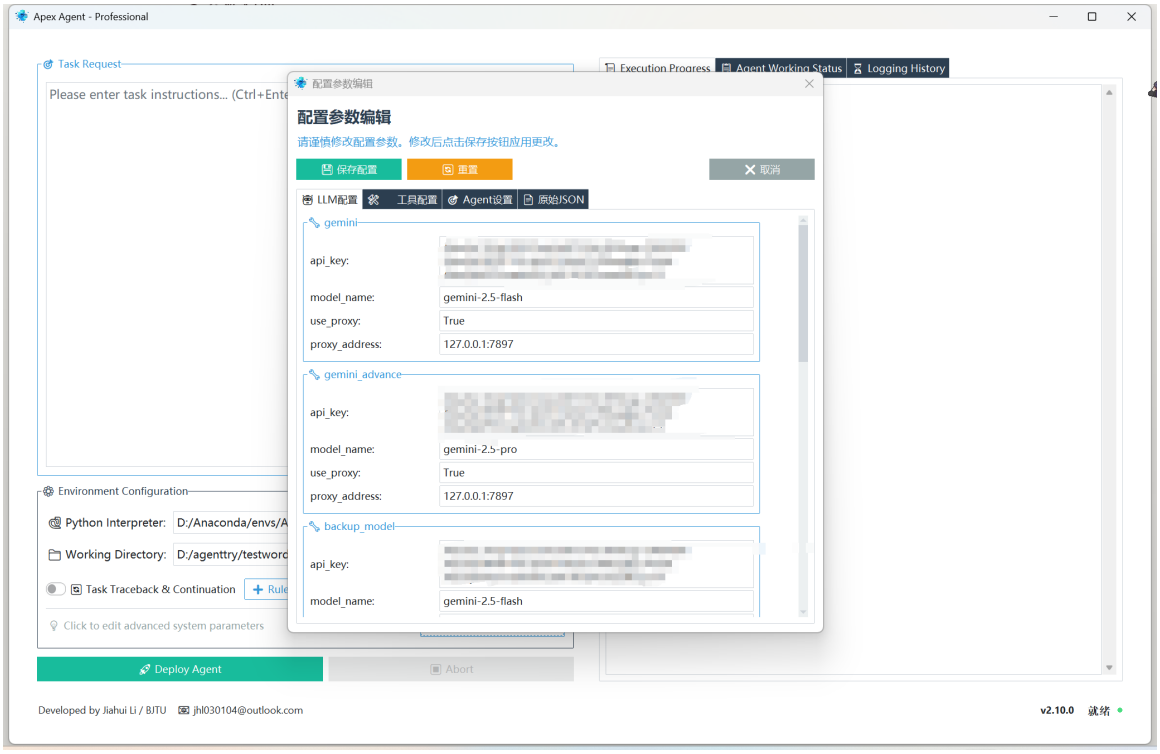
运行`ApexAgent.msi`安装包，快速下载。

2.2 初始参数配置

首次使用需要配置智能体参数，点击主界面的`Advanced Configuration`按钮进行参数初始配置：



- **LLM配置：**包括Gemini LLM API、第三方中转LLM API（可选，不填可以正常使用但代理无法访问到外部LLM寻求帮助）；

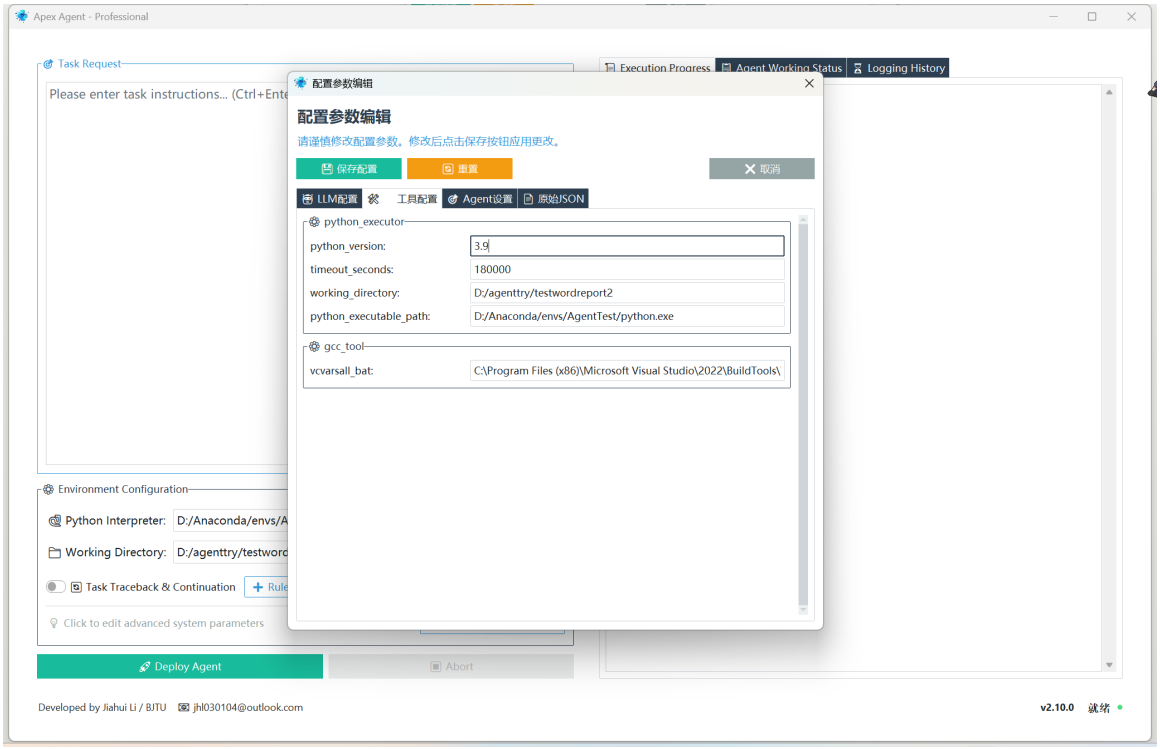


配置Gemini API时，由于系统是轮询方式，因此可以填多个API-KEY，每一行填一个API-KEY，**不要添加分号或逗号隔开**，示例填写如下，注意不要留有空白行，否则会被当成空API-KEY去轮询，会导致报错：

```
MYAPIKEY1
MYAPIKEY1
MYAPIKEY1
MYAPIKEY1
......
```

此外，由于Gemini不支持部分地区用户使用，因此国内需要能够科学上网，并打开**Tun虚拟卡模式（极其重要）**，不要使用系统代理模式，否则部分功能将会受限导致应用卡住。无需科学上网即可使用地区的用户，可以直接在use_proxy位置填false即可。

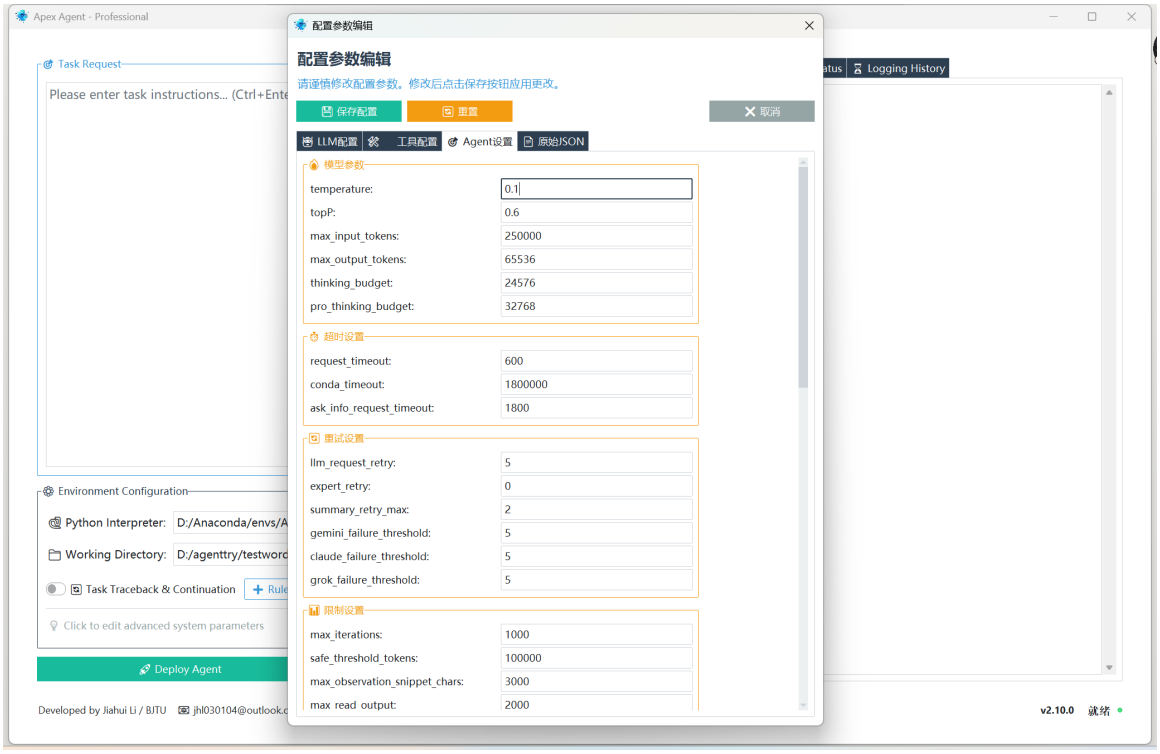
- **工具配置**：需要配置给代理使用的python解释器路径和c/c++生成工具里面的vcvarsall.bat路径，python解释器也可以在主界面直接切换和配置，其余可以暂时默认；



工具配置中，python_version参数已经弃用，默认即可，无需改动（作者太懒不想动代码）；

timeout_seconds代表执行python代码和CMD命令最长的超时时间，一般按照默认即可（180000s=50h，一般程序都能执行完毕）

- **Agent设置**：这里面的参数为调节智能体参数，如温度、topP等，一般情况下都选择默认即可（作者已经调好的，一般情况下都可以正常使用的参数），由于参数太多，如需了解其余参数的详情，欢迎咨询作者（jhl030104@outlook.com）。



- **原始JSON**：高级用户（即能看明白的）可以直接编辑JSON配置。请确保JSON格式正确即可。（本质上前面3栏只是对**原始JSON**进行了分组分栏，修改哪一个都可以）

修改完成后，点击保存配置即可，如果过程中想要退回未修改前的，点击重置按钮即可（已经保存过的无法重置返回，因为已经写入配置文件当中了）

2.3 使用方法

下面截图是有关主界面的详细介绍，请参考。


Apex Agent - Professional
Task Request
Please enter task instructions... (Ctrl+Enter / click Deploy Agent to send)
该区域内输入你的任务指令
Execution Progress
Agent Working Status
Logging History
实时日志
工作状态文档（相当于todolist）
历史日志
Environment Configuration
Python Interpreter: D:/Anaconda/envs/AgentTest/python.exe
Browse
Working Directory: D:/agenttry/testwordreport2
Browse
可以快速配置python解释器路径和当前任务执行的工作目录
Task Traceback & Continuation
Rule Addition
Role Card: GeneralRole
模式选择，第一个是任务接续（用于任务中断时可以继续接着执行），
第二个是添加全局规则；第三个是角色卡，用于约束Agent在垂域表现
Deploy Agent
Abort
发送任务给Agent执行
中断执行（非强制性中断，而是给Agent发送中止信号，
当执行完当前一步工作后才会起效）
v2.10.0
就绪

好的，请开始使用吧~