我们仍然需要 GitHub 搜索框的情况

```yaml
jobs:
  awesome-cv-copy:
    runs-on: ubuntu-latest
    steps:

     # ...

      - name: 安装 TeX Live 2023
        if: steps.cache-texlive.outputs.cache-hit != 'true'
        run: |
          # 安装 TeX Live 安装程序的依赖
          sudo apt-get update
          sudo apt-get install -y perl wget xz-utils

          # 下载 TeX Live 安装程序
          wget http://mirror.ctan.org/systems/texlive/tlnet/install-tl-unx.tar.gz
          tar -xzf install-tl-unx.tar.gz
          cd install-tl-*/

      # ...

      - name: 安装缺失的 LaTeX 包
        run: |
          sudo /usr/local/texlive/2023/bin/x86_64-linux/tlmgr install etoolbox adjustbox

      - name: 确认包安装
        run: |
          kpsewhich etoolbox.sty
          kpsewhich adjustbox.sty

      - name: 运行 make awesome-cv-copy
        run: make awesome-cv-copy
```

我正在处理上面的 GitHub Actions 脚本。

我需要在 GitHub 上搜索 `etoolbox adjustbox language:YAML` 代码的确切匹配。

我遇到以下错误：

```
2025-01-07T22:34:58.6493408Z
2025-01-07T22:34:58.6493741Z ! LaTeX Error: File adjustbox.sty' not found.
2025-01-07T22:34:58.6494172Z
2025-01-07T22:34:58.6494593Z Type X to quit or <RETURN> to proceed,
2025-01-07T22:34:58.6495322Z or enter new name. (Default extension: sty)
```

我正在特别搜索 `etoolbox adjustbox language:YAML`，而 GitHub 中的结果有限，只有 53 个 YAML 文件包含 `etoolbox` 和 `adjustbox`。我需要 **精确匹配**。

尽管我们已经进入了大语言模型的时代，但精确匹配的搜索仍然至关重要。这对于检查某个内容的确切含义或找到准确的工作代码特别重要。同样，像 Google、Twitter 或其他平台也依赖于精确搜索来理解含义。我们不希望看到 AI 生成的结果或有细微错误的结果。

为了训练大语言模型，我们可以开发一个找到精确匹配的系统。也许我们可以结合 **KMP（Knuth-Morris-Pratt）**搜索算法与 **transformer 架构** 来增强搜索能力。结合 KMP 和 Transformer 可以帮助更准确地找到特定代码的搜索结果。

目前，大语言模型无法按文件语言（如 YAML 或 Python）进行筛选。然而，现实世界中的大量信息是以这种方式组织的。这意味着我们可以通过文件来训练大语言模型。如果我们按文件类型组织所有文本数据，就可以更好地训练模型。因此，对于大语言模型，我们需要预定义文件语言。默认情况下，它可以是 “text”，但我们也可以像 GitHub 搜索一样定义其他语言。这样，搜索结果就会像 GitHub 搜索结果一样返回文件。

重要的是 **文件格式** 或 **扩展名**，而不是文件名。以下是一些示例：

> Python, JavaScript, Java, Ruby, Go, C++, C, C#, TypeScript, HTML, CSS, PHP, Swift, Kotlin, Rust, Objective-C, Bash, Markdown, R, Lua, Haskell, MATLAB, Perl, SQL, Dockerfile, YAML, JSON, TOML, VHDL, TeX, LaTeX, Assembly, GraphQL
>
> .py, .js, .java, .rb, .go, .cpp, .cc, .cxx, .h, .c, .cs, .ts, .html, .htm, .css, .php, .swift, .kt, .kts, .rs, .m, .h, .sh, .md, .r, .lua, .hs, .m, .pl, .pm, .sql, Dockerfile, .yaml, .yml, .json, .toml, .vhdl, .vhd, .tex, .asm, .graphql, .gql

然而，当用户的提示混合了普通文本和类似文件的表达式和符号时，就很难进行这样的搜索。例如，在像 Stack Overflow 这样的平台上，问题或答案往往包含混合了文本和代码片段或文件表达式。

但显然，在这个领域中，我们可以设想一些新产品来弥合自然语言搜索和基于文件的搜索之间的差距。