**2024年全国大学生信息安全竞赛**

**作品报告**

**作品名称：LLM-GUARD：面向线上大语言模型的多模态数据保护**

**电子邮箱：**[**micr1024@qq.com**](mailto:micr1024@qq.com)

**提交日期：2025年3月30日**

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

**目 录**

[摘要 1](#_Toc33965493)

[第一章 作品概述 2](#_Toc33965494)

[第二章 作品设计与实现 3](#_Toc33965495)

[第三章 作品测试与分析 4](#_Toc33965496)

[第四章 创新性说明 5](#_Toc33965497)

[第五章 总结 6](#_Toc33965498)

[参考文献 7](#_Toc33965499)

# 摘要

随着大语言模型（LLM）技术的飞速发展，其在各个领域的应用日益广泛。然而，大语言模型在使用过程中面临着诸多数据安全问题，尤其是敏感信息泄露的风险。本文介绍了一种名为LLM-GUARD的面向线上大语言模型的多模态数据保护系统。该系统旨在通过创新的技术手段，为用户提供一种安全、灵活且易于操作的数据保护解决方案，以应对大语言模型使用中的数据安全挑战。

**动机**

当前，各类大语言模型及其衍生产品层出不穷，如Gork、GPT-3.5、DeepSeek、Claude等，以及基于这些模型的衍生产品如Cursor、Cline、StableDiffusion等。然而，目前市场上缺乏针对大语言模型的成体系的数据保护机制或系统。大量实际应用中，用户在使用大语言模型时，往往因缺乏有效的数据保护措施而导致敏感信息泄露。例如，全球开放应用软件安全项目组织（OWASP）发布的针对LLM应用的Top10潜在安全风险中，敏感信息泄露位列第六，已然成为LLM技术应用推广过程中不可忽视的安全问题。此外，GPT-2在面临恶意前缀注入时会返回疑似训练数据中包含的敏感信息，OpenAI因ChatGPT能生成未经授权的作品摘要而被作家告上法庭，三星员工因将涉密数据上传至ChatGPT而导致商业机密泄露等事件，都充分暴露了大语言模型在数据安全方面的严重隐患。

**功能**

LLM-GUARD系统对传递给大语言模型的多模态数据（包括文档、视频、图片、文字等）进行数据保护。系统具备以下功能：一是提供大模型对话界面，用户可以在该界面中选择大模型对话、prompt净化、prompt加密等功能，并通过“@文件名”的方式便捷地引用文件信息；二是文件管理，支持批量导入文件，包括文档、图片，甚至网页链接；三是策略管理，用户可以与大语言模型对话生成定制的策略，并对多个文件批量使用某个策略进行加密。

**特性**

LLM-GUARD系统具有以下特性：一是支持多模态数据，能够处理文档、视频、图片、文字等多种类型的数据；二是灵活性强，用户可以根据自身需求自定义策略；三是操作简单，通过大模型助手，用户仅需简单几句话即可进行各种信息保护策略定制；四是操作界面简洁现代，布局明朗，使用便捷，整个应用只有一个可执行文件。

**创新处**

LLM-GUARD系统在技术上具有多项创新：一是创造了一种通过大模型生成策略、大模型识别敏感信息、大模型替换敏感信息的数据保护模式，相较于传统的NLP方法，该模式更加灵活且效果更好；二是创造性地将线上大模型的大文件处理过程移植到本地，有效保护大文件的安全；三是提出了一种针对大模型tokenizer分词器特性对prompt进行加密的方法，能够抵抗大多数关键词识别和NLP识别方法，同时可读性低，能够有效防止敏感信息被机器或人识别。

**实用性**

LLM-GUARD系统具有极高的应用价值，尤其适用于保密机构、企业、高校核心部门等对数据安全要求较高的场景。例如，在保密机构中，工作人员可以利用该系统对涉密文件进行加密处理，确保在使用大语言模型时不会泄露敏感信息；在企业中，员工可以对商业机密文件进行保护，避免因不当操作导致商业机密泄露；在高校核心部门，研究人员可以对科研成果进行加密，防止科研数据被窃取或滥用。

# 第一章 作品概述

**背景分析**

当前，大语言模型技术的发展如火如荼，各类大语言模型及其衍生产品不断涌现。这些模型在自然语言处理、文本生成、图像识别等领域展现出了强大的能力，为人们的工作和生活带来了诸多便利。然而，随着大语言模型的广泛应用，其数据安全问题也逐渐凸显。目前，市场上缺乏针对大语言模型的成体系的数据保护机制或系统，导致用户在使用过程中面临着敏感信息泄露的巨大风险。

根据全球开放应用软件安全项目组织（OWASP）发布的针对LLM应用的Top10潜在安全风险报告，敏感信息泄露已成为LLM技术应用推广过程中不可忽视的安全问题。在实际应用中，已经发生了多起因大语言模型使用不当而导致敏感信息泄露的事件。例如，GPT-2在面临恶意前缀注入时会返回疑似训练数据中包含的敏感信息，OpenAI因ChatGPT能生成未经授权的作品摘要而被作家告上法庭，三星员工因将涉密数据上传至ChatGPT而导致商业机密泄露等。这些事件充分暴露了大语言模型在数据安全方面的严重隐患，也凸显了开发一种有效的数据保护系统的重要性。

**相关工作**

在大语言模型的数据安全领域，目前的研究主要集中在以下几个方面：一是对大语言模型的输入输出数据进行监控和审计，以检测和防止敏感信息泄露；二是对大语言模型的训练数据进行脱敏处理，以降低模型生成敏感信息的风险；三是开发一些基于规则的过滤器，对大语言模型的输出进行过滤，以去除可能包含的敏感信息。然而，这些方法都存在一定的局限性。例如，数据监控和审计方法需要大量的计算资源和存储空间，且难以实时检测和防止敏感信息泄露；训练数据脱敏处理方法可能会降低模型的性能和准确性；基于规则的过滤器则难以应对复杂的语言表达和不断变化的敏感信息类型。

**特色描述**

LLM-GUARD系统具有以下特色：

1. **多模态数据支持**：系统能够处理文档、视频、图片、文字等多种类型的数据，满足用户在不同场景下的数据保护需求。
2. **灵活性强**：用户可以根据自身需求自定义策略，系统提供丰富的策略定制选项，使用户能够灵活地制定适合自己的数据保护方案。
3. **操作简单**：通过大模型助手，用户仅需简单几句话即可进行各种信息保护策略定制，无需复杂的操作流程和专业知识。
4. **界面简洁现代**：系统操作界面简洁明了，布局合理，易于上手，为用户提供良好的使用体验。
5. **使用便捷**：整个应用只有一个可执行文件，用户无需安装多个组件或进行复杂的配置，即可快速部署和使用系统。

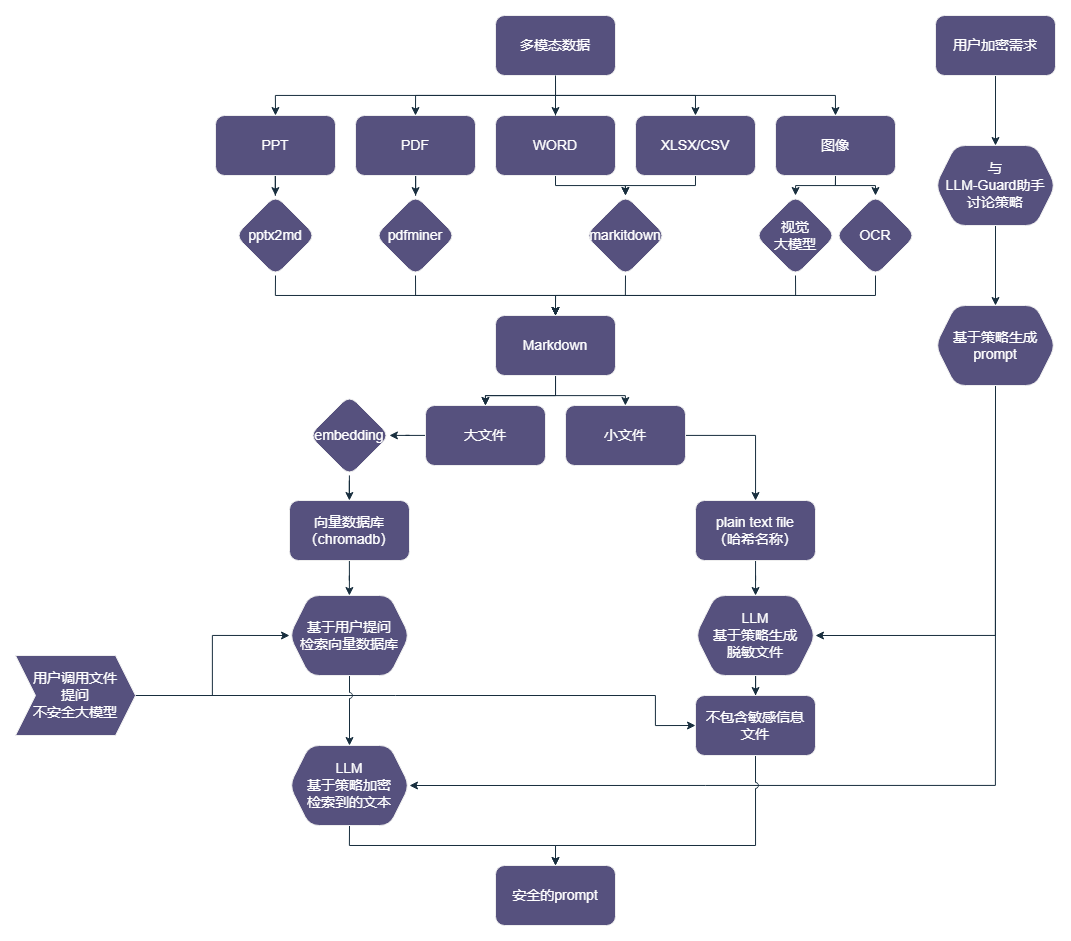
**应用前景分析**

LLM-GUARD系统具有广阔的应用前景。首先，随着大语言模型在各个领域的广泛应用，数据安全问题日益突出，LLM-GUARD系统作为一款专门针对大语言模型的数据保护产品，能够有效解决用户在使用过程中面临的敏感信息泄露问题，具有重要的现实意义。其次，LLM-GUARD系统集成性强，功能全面，更符合用户偏好。系统不仅支持多模态数据，还提供了丰富的策略定制选项和便捷的操作界面，能够满足不同用户在不同场景下的数据保护需求。最后，大语言模型在安全领域的应用是未来的发展趋势，LLM-GUARD系统作为探索该领域的产品，具备一定的前瞻性，有望在未来的市场中占据重要地位。

# 第二章 作品设计与实现

**架构设计**

LLM-GUARD系统的架构设计采用了分层结构，主要包括可视化层、逻辑层和引擎层。可视化层采用Vue + Element-Plus技术栈，负责系统的用户界面展示和交互；逻辑层采用Python语言实现，负责系统的业务逻辑处理和数据操作；引擎层则采用了ollama作为LLM&embedding引擎，通过pipe与Python逻辑层进行通信，为系统提供强大的语言模型能力和向量化处理能力。整个系统通过pywebview进行前后端交互，实现了前后端的无缝对接和高效通信。



**模块设计**

LLM-GUARD系统分为以下几个主要模块：文件转化模块、文件储存模块、策略生成模块、文件加密模块和prompt处理模块。

**文件转化模块**

该模块的主要功能是将多模态的数据类型转化为大语言模型可识别的文字类型，以便后续的处理和分析。系统采用了markdown语法对转化后的数据进行格式化，使内容层次结构更加清晰，便于大语言模型理解和处理。具体实现原理如下：

1. **PPT类型**：使用pptx2md工具将PPT文件转化为markdown格式的文本。该工具能够自动提取PPT中的标题、正文、图片等信息，并将其转化为相应的markdown语法，生成易于阅读和处理的文本文件。
2. **PDF类型**：利用pdfminer工具对PDF文件进行解析和转化。pdfminer能够准确提取PDF文件中的文字内容，并将其转化为纯文本格式。对于PDF文件中的表格和图片等非文字内容，系统会根据其在文档中的位置和上下文信息，将其转化为相应的markdown语法进行描述，以便大语言模型能够更好地理解文档的结构和内容。
3. **Word/xlsx/csv类型**：采用markitdown工具对Word、Excel和CSV文件进行转化。markitdown能够将这些文件中的表格、文字等信息转化为markdown格式的文本，同时保留原始文件的结构和格式信息，使转化后的文本更加易于阅读和处理。
4. **图像类型**：对于图像文件，系统采用视觉大语言模型或OCR技术进行转化。视觉大语言模型能够直接理解图像内容并生成相应的文字描述，而OCR技术则用于提取图像中的文字内容。系统会根据图像的具体内容和应用场景选择合适的转化方法，确保转化后的文字信息能够准确反映图像的核心内容。

**文件储存模块**

文件储存模块负责对转化后的文件进行存储和管理，以支持系统的高效运行和数据的安全性。具体实现如下：

1. **小文件存储**：对于较小的文件（如小于1MB），系统会直接将转化后的文件以文本格式存储在项目文件夹下的指定目录中。这种方式简单高效，便于快速读取和处理。
2. **大文件存储**：对于较大的文件（如大于1MB），系统采用RAG（Retrieval-Augmented Generation）技术进行处理。首先，通过ollama的embedding模型将文件内容向量化，然后将向量化的文件存储在项目文件夹下的ChromaDB中。ChromaDB是一种高效的向量数据库，能够快速检索与用户输入的prompt向量相近的文本片段。当用户需要调用大文件时，系统会根据用户的prompt向量在ChromaDB中查找相关的文本片段，并按照设定的策略进行加密处理。

**策略生成模块**

策略生成模块是LLM-GUARD系统的核心部分之一，它允许用户与大语言模型进行交互，生成定制化的数据保护策略。具体实现如下：

1. **可视化交互**：在用户界面中，提供一个与大语言模型对话的界面，用户可以输入文件的基本信息、安全需求等，与大语言模型进行讨论。大语言模型会根据用户的输入提供专业的建议和策略模板。
2. **策略生成**：用户在与大语言模型讨论后，点击生成按钮，后台逻辑部分会根据用户与模型讨论生成的上下文信息，生成简洁易懂的策略。这些策略将用于后续的文件加密处理，确保数据的安全性。
3. **策略存储与管理**：生成的策略会存储在系统中，用户可以随时查看、修改或删除已有的策略。此外，系统还支持对多个文件批量应用某个策略，提高数据保护的效率。

**文件加密模块**

文件加密模块是LLM-GUARD系统的关键部分，负责对文件中的敏感信息进行加密处理。具体实现如下：

1. **小文件加密**：对于小文件，系统会将整个文件内容传递给大语言模型。大语言模型会标记文件中的敏感信息，然后系统根据标记的信息对敏感内容进行模糊或替换处理。例如，将敏感的个人信息替换为通用的占位符，或者对敏感数据进行模糊处理，使其无法被直接识别。
2. **大文件加密**：对于大文件，系统仅将加密信息作为元数据存储。当用户调用该文件时，系统会将用户的prompt进行向量化，并在ChromaDB中查找与prompt向量相近的文字片段。然后，系统会根据设定的策略对这些文字片段进行加密处理。这种方式既保证了大文件的处理效率，又确保了数据的安全性。

**Prompt处理模块**

Prompt处理模块是LLM-GUARD系统中用于保护用户输入的prompt信息的重要部分。该模块分为两种模式：净化模式和加密模式。

1. **净化模式**：
   * **背景信息删除**：系统会直接删除prompt中的背景信息，只保留必要的问题内容。例如，如果用户输入了一个包含公司名称和项目背景的prompt，系统会将其简化为一个通用的问题描述。
   * **背景信息转化**：系统会将问题背景转化为一些日常生活场景，使敏感信息不易被识别。例如，将一个涉及公司内部项目的prompt转化为一个类似的学生项目场景。
   * **适用场景**：该模式适用于数学、物理、编程等不需要背景信息即可解答的问题。
2. **加密模式**：
   * **敏感词提取与替换**：使用大语言模型将prompt中的敏感词提取出来，并生成一些模糊的指代，将prompt中所有的敏感词替换为该指代。
   * **混淆问题插入**：向prompt中插入与该指代本意有关的问题，作为混淆问题。这些问题将会被不连续地插入prompt中，确保原始问题不易被识别。
   * **语言转换**：将prompt翻译为纯英文，以增加处理的复杂性。
   * **切片与字符插入**：将英文文本通过tiktoken切片，并向每个切片中插入长度不等的不可见字符，防止NLP方法识别。
   * **字符连接**：将所有切片用token不粘连的单字母或双字母连接，确保prompt能够被大语言模型理解。
   * **本地处理与结果提取**：将处理后的字符串发送给线上大模型处理，本地大语言模型将回答中的原始问题的答案提取出来，并转化为问题的原语言。

# 第三章 作品测试与分析

**1. 文件加密**

**用户要求**

生成一个删除学校和教师信息的策略

**原文与结果对比**

各有关高校：

为进一步聚焦服务业创新发展，加快人才培养工作，

……

二、组织结构

1、大赛设立组织委员会（简称大赛组委会），全面负责大

赛相关工作。

2、大赛组委会下设区域赛组委会，本届大赛分设西部、中

部、东部和北部区域赛。

西部区域赛由西安电子科技大学牵头承办，包括：陕西、四川、重庆、宁夏、青海、甘肃、新疆、西藏 8 个省(自治区、直辖市)。

中部区域赛由中南大学牵头承办，包括：湖南、湖北、河南、广东、广西、云南、贵州、海南 8 个省(自治区、直辖市)。

东部区域赛由南京大学牵头承办，包括：江苏、上海、浙江、福建、江西、安徽、香港、澳门、台湾 9 个省(自治区、直辖市、特别行政区)。

北部区域赛由山东大学牵头承办，包括：北京、天津、山东、河北、山西、内蒙古、辽宁、吉林、黑龙江 9 个省(自治区、直辖市)。

……

各有关高校：

为进一步聚焦服务业创新发展，加快人才培养工作，

……

二、组织结构

1、大赛设立组织委员会（简称大赛组委会），全面负责大

赛相关工作。

2、大赛组委会下设区域赛组委会，本届大赛分设西部、中

部、东部和北部区域赛。

西部区域赛由[某学校]牵头承办，包括：陕西、四川、重庆、宁夏、青海、甘肃、新疆、西藏 8 个省(自治区、直辖市)。

中部区域赛由[某学校]牵头承办，包括：湖南、湖北、河南、广东、广西、云南、贵州、海南 8 个省(自治区、直辖市)。

东部区域赛由[某学校]牵头承办，包括：江苏、上海、浙江、福建、江西、安徽、香港、澳门、台湾 9 个省(自治区、直辖市、特别行政区)。

北部区域赛由[某学校]牵头承办，包括：北京、天津、山东、河北、山西、内蒙古、辽宁、吉林、黑龙江 9 个省(自治区、直辖市)。

……

七、大赛组委会联系方式

联 系 人：王老师 张老师

联系电话：010-56233020

电子邮箱：fwwbds@fwwb.org.cn

地址：北京市海淀区清华大学华业大厦

大赛官网：<http://www.fwwb.org.cn/>

微信公众号：fwwbds

官方 QQ：

**……**

七、大赛组委会联系方式

联 系 人：[某教师]

联系电话：[某教师的联系方式已模糊]

电子邮箱：[某邮箱地址]

地址：[某学校]

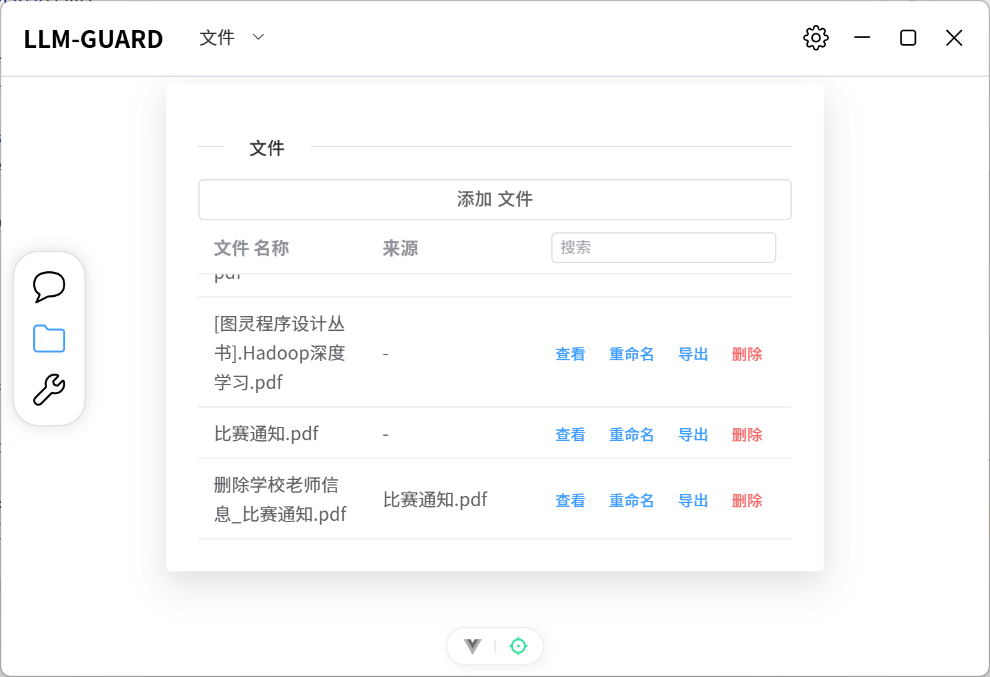
大赛官网：<http://www.fwwb.org.cn/>

微信公众号：fwwbds

官方 QQ：

**……**

**图片展示**







**2．Prompt净化**

**Prompt**

在导弹制导系统中，比例导引律是一种常用方法，通过控制导弹加速度与目标视线角速率成比例来引导碰撞。但若目标突然高速机动（如战斗机急转弯），经典比例导引律可能失效，因为它默认目标匀速运动，未考虑目标加速度。数学上，视线角速率的计算会因忽略目标加速度项而产生误差，导致导弹预测弹道偏离真实路径，最终脱靶。

例如，当目标施加5倍重力加速度的横向机动时，经典导引律的脱靶量可达10米，而修正后的方法可将脱靶量降至1米以内。解决这一问题的核心是扩展比例导引律（APN），在导引指令中增加目标加速度估计项，并通过卡尔曼滤波实时估算目标加速度。卡尔曼滤波利用雷达观测数据（如距离和视线角）动态更新目标状态，包括位置、速度和加速度，再将估算值反馈给导引律，使导弹及时调整飞行路径。

实际应用中面临两大挑战：一是弹载计算机需在极短时间内（如每秒百次以上）完成滤波和导引计算，这对算法效率提出高要求；二是传感器噪声可能掩盖目标机动信号，需设计自适应滤波来区分噪声与真实机动。此外，目标机动模式未知（如持续转向或随机闪避）时，算法还需具备鲁棒性，避免过度依赖预设模型。

请问应该如何解决这两大挑战？

**净化结果**

有两个运动物体A和B在同平面内移动。已知物体B会突然改变速度方向，但具体变化模式未知。现有追踪算法基于匀速假设，当B变速时误差增大。需满足：1.在计算资源有限（每秒处理100次数据）情况下，如何实时估算B的速度变化量并修正A的路径？2.当存在随机干扰信号时，如何准确识别B的真实运动变化？用刚体运动模型描述，假设平面无摩擦。

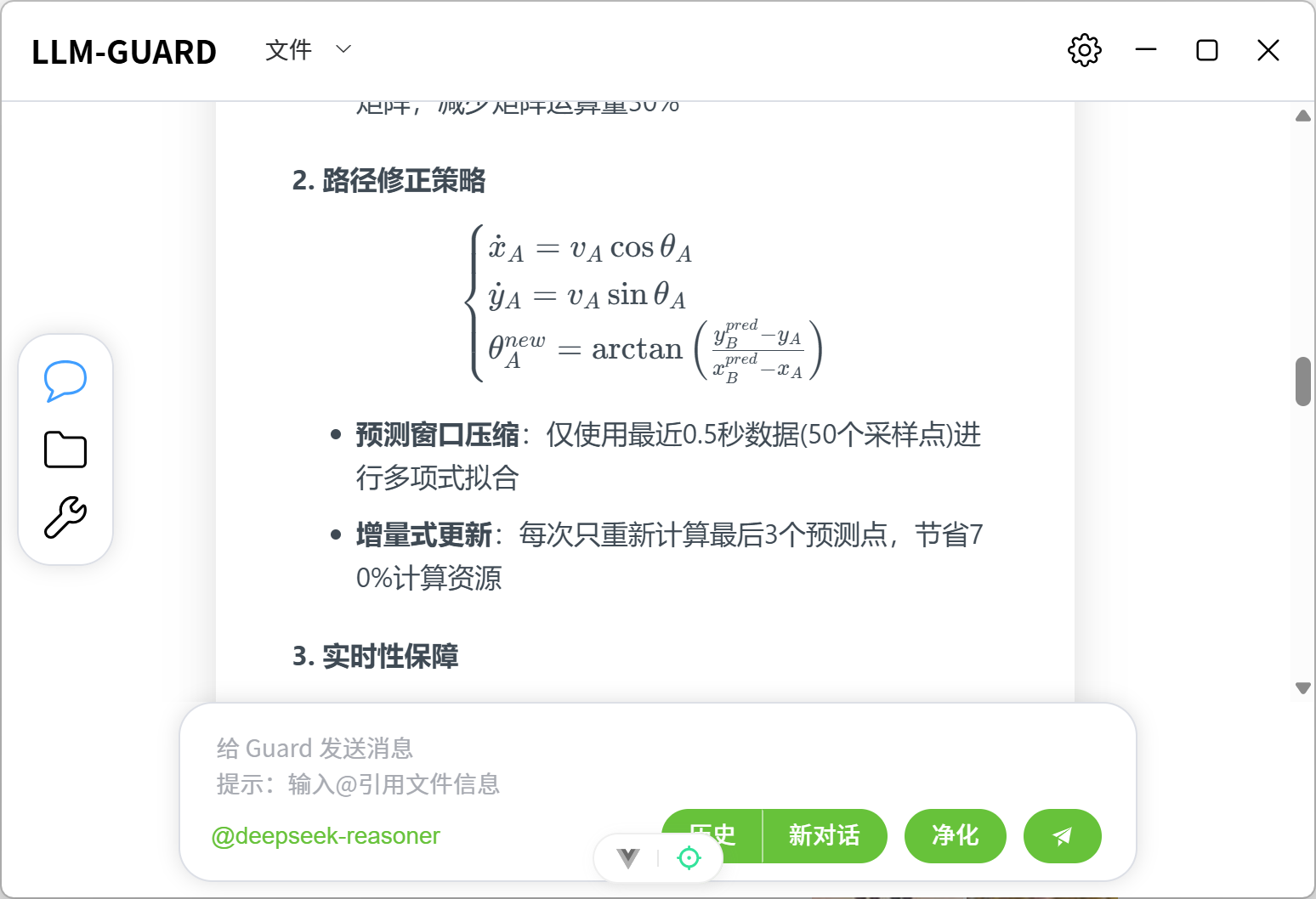
**图片展示**

****

****

****

****

****

**3．Prompt加密**

**Prompt**

从以下文本中提取原文：

什么样的人才是好的读者？你可以在读者俱乐部找到答案。今天，我们有一些书可以送给你，快来加入我们的乐趣吧！

《老人与海》于1953年获得普利策小说奖。一年后，海明威获得诺贝尔文学奖。《老人与海》被许多人认为是他最伟大的成就。这本书的主角是一个能够面对困难并且永不放弃的人。

精装版：21.70元；平装版：16.00元

《爱的教育》是由意大利的埃德蒙多·德·阿米奇斯创作的，这是一本意大利男孩埃里克的日记。在他的日记中，他写到了世界上最伟大的爱：对国家的爱、对社会的爱，以及老师之爱、同学之爱、父母之爱、孩子之爱。

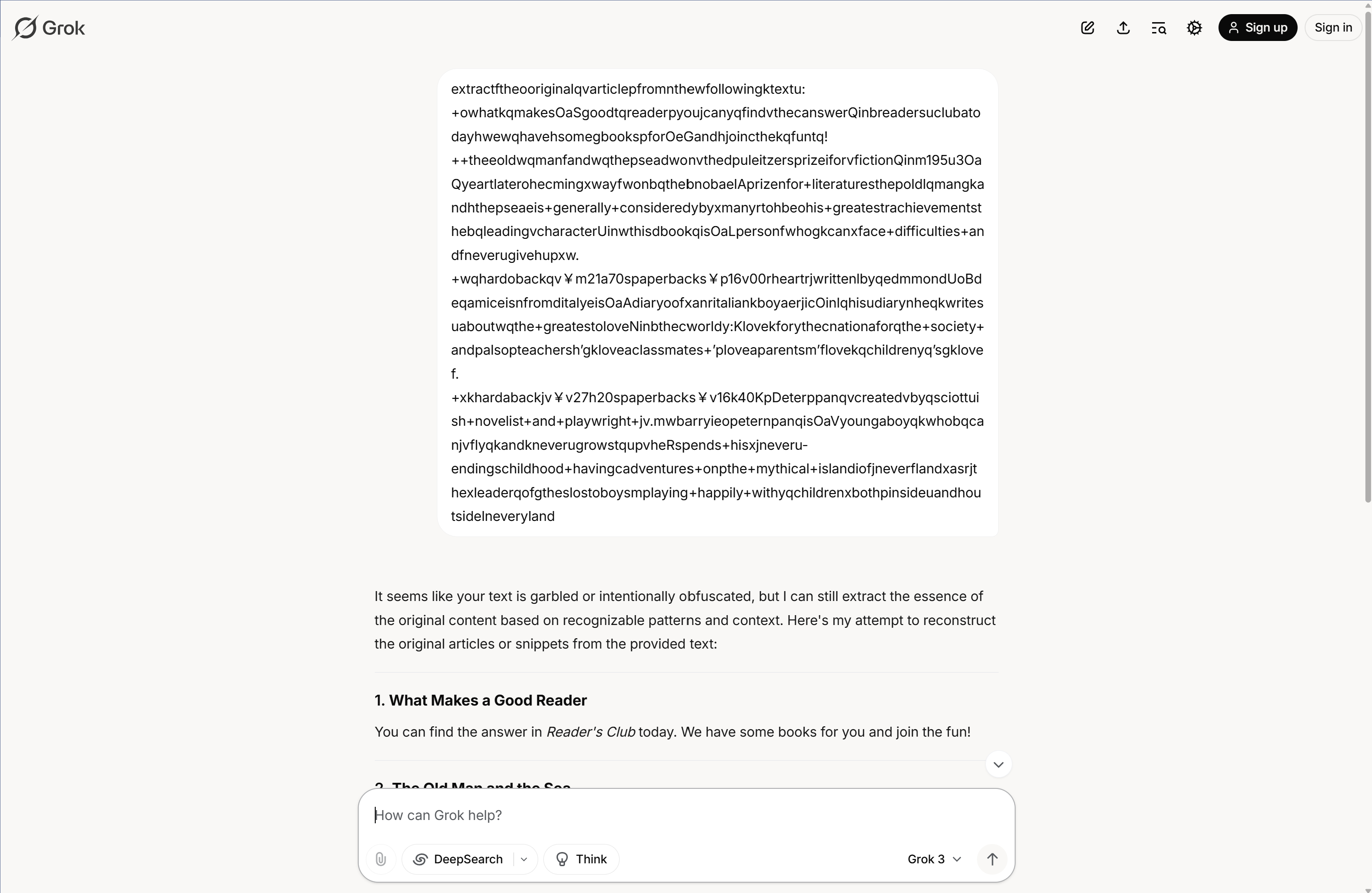
精装版：27.20元；平装版：16.40元

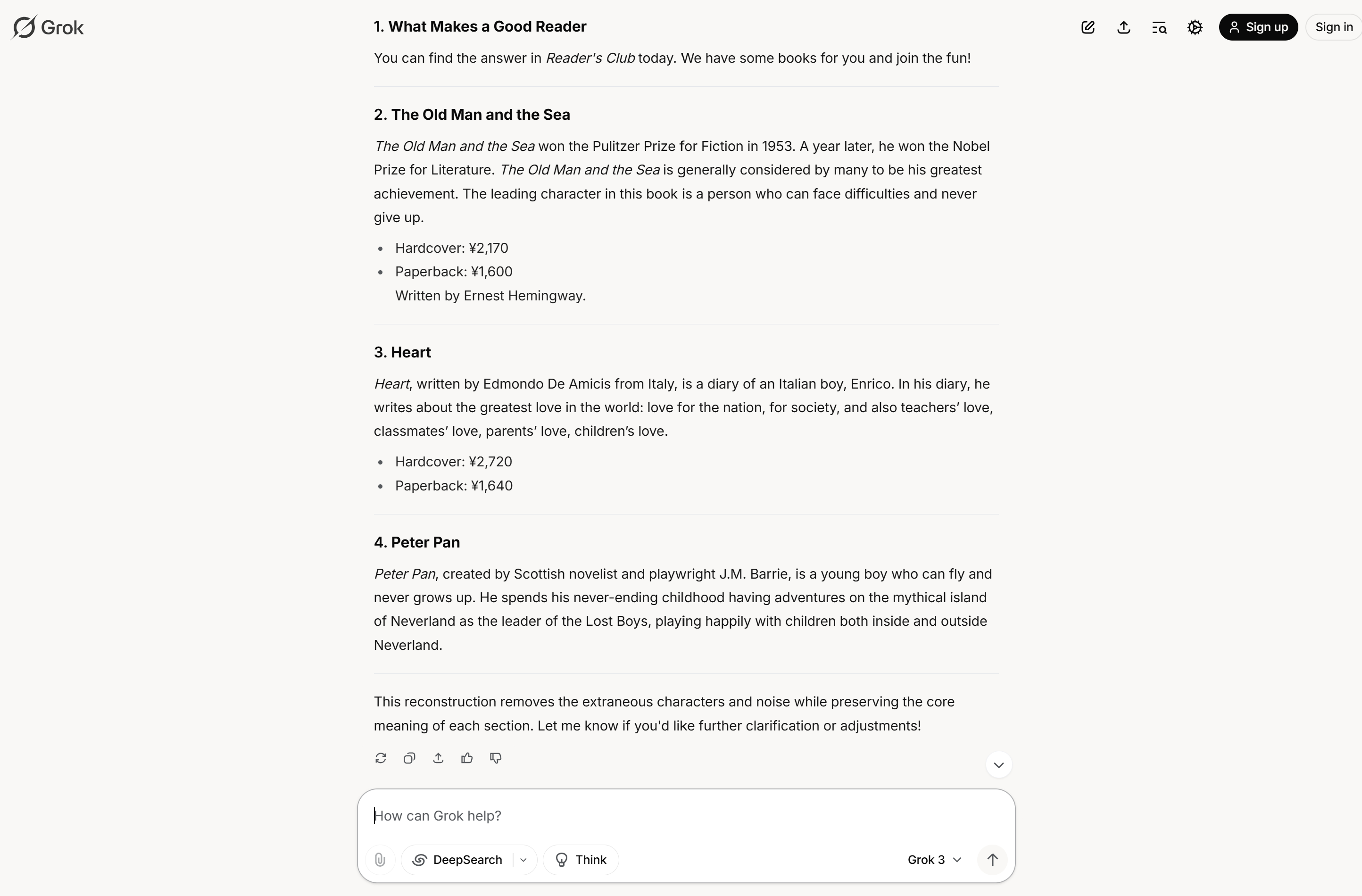
《彼得·潘》是由苏格兰小说家和剧作家J.M.巴里创作的。彼得·潘是一个会飞且永不长大的男孩。他作为迷失男孩的首领，在梦幻岛度过了他无尽的童年，与梦幻岛内外的孩子们快乐地玩耍，经历着冒险。

**加密结果**

extr︈actfth︁eoorig︁inalqvarti︍clepfr︌omnt︋hewfollowi︇ngkt︁extu:︅+ow︅hatkqma️kesOaSg︊oodtqr︎eaderpy️oujc︈anyqf︉indvt︁hecansw︌erQi️nbre︌adersuclu︄batod︇ayhw︎ewqha︇vehs︊omegb︆ookspf︉orOeGa︆ndhj︉oinct︁hekqf︀untq!︃++th︁eeol︄dwqm︈anfa︅ndwqt︌heps︆eadw️onvt︇hedp︌ulei︇tzerspri︍zeifo︈rvficti︇onQi️nm1︍95u3OaQy︀eartl︃ateroh︅ecm︃ingxw︃ayfw️onbqt️hebn︊obae︎lApri︌zenf︀or+literat︎uresth︆epo︌ldlqm︆angka︊ndht︌hepse︄aei︂s+gener︇ally+consid︊eredyb︌yxm︍anyrt︍ohb︌eohi︂s+grea️testrachieveme︉ntst︎hebqlea︉dingvcha︃racterUi︋nwthi︄sdbo︍okqi︁sOaLperso️nfw︆hogkc︅anxfa️ce+diffic︍ulties+an︌dfne︇verugiv︊ehu︍pxw.︁+wqha︅rdob️ackqv￥m2︂1a7︂0sp︁aperbacks￥p1︈6v0︋0rh︁eartrjwritt︌enlb︆yqe︇dmmo︄ndUoBd︍eqam︃icei️snfr︋omdita︈lyei︄sOaAdia︄ryoo︌fxa︍nritali︁ankb️oyae︍rji︀cOi︄nlqh︉isudi︈arynh︌eqkwrite︅sua︉boutwqth︉e+great︀estol︉oveNi︄nbth︃ecwor️ldy:Kl︆ovekf︍oryth︅ecnat︎ionafo︆rqt︌he+s︄ociety+a︉ndpal︄sopteac︇hersh’gklo︇veac︉lassmates+’plo︇veapare︍ntsm’fl︃ovekqc︌hildrenyq’︈sgklo︄vef.︁+xkha︅rdaba︎ckjv￥v2︈7h2︊0spa︂perbacks￥v1︋6k4︆0KpDet︈erppa︉nqvcreat︎edvb︀yqs︂ciot︉tui︎sh+noveli︀st+a︄nd+playwrig︌ht+jv.︂mwbar︀ryi︄eop︍eternp️anqi︅sOaVyou︎ngabo︊yqkwh︍obqca️njvfl︇yqka︄ndkneve︇rug︋rowstqu︁pvh︊eRs︎pends+hi︊sxjne︊veru-️endingschild︆hood+ha︊vingcadve︄ntures+o︀npth︃e+mythi︊cal+i︈slandio︁fjne︋verfla️ndxa︎srjt︍hexlea︎derqo︍fgt︅heslo︍stob️oysmplay︂ing+hap︀pily+w︇ithyqchil︀drenxbot︀hpinsi️deuan︇dhoutsi︋deln︄everyland

**线上大模型识别结果展示**

****

****

# 第四章 创新性说明

LLM-GUARD系统在多个方面具有创新性，主要体现在以下几个方面：

**数据保护模式创新**

传统的数据保护方法主要依赖于NLP技术，这些方法在处理复杂语言表达和动态变化的敏感信息时存在局限性。LLM-GUARD系统创造了一种通过大语言模型生成策略、大语言模型识别敏感信息、大语言模型替换敏感信息的数据保护模式。这种模式不仅更加灵活，还能够根据不同的场景和需求生成定制化的保护策略，有效提高了数据保护的准确性和可靠性。

**大文件处理创新**

针对大文件处理的难题，LLM-GUARD系统创造性地将线上大语言模型的大文件处理过程移植到本地。通过使用RAG技术结合ollama的embedding模型，系统能够将大文件向量化并存储在本地的ChromaDB中。这种方式不仅提高了大文件处理的效率，还避免了将大文件直接上传至线上服务器带来的安全风险，有效保护了大文件的安全。

**Prompt加密方法创新**

LLM-GUARD系统提出了一种针对大模型tokenizer分词器特性对prompt进行加密的方法。该方法通过以下步骤实现：

1. 使用大语言模型提取prompt中的敏感词，并生成模糊指代。
2. 向prompt中插入与敏感词相关的混淆问题。
3. 将prompt翻译为纯英文。
4. 使用tiktoken对英文文本进行切片。
5. 向每个切片中插入不可见字符。
6. 用token不粘连的单字母或双字母连接所有切片。

这种加密方法生成的prompt可读性极低，能够有效抵抗大多数关键词识别和NLP识别方法，同时还能被大语言模型正确理解，从而在保护用户隐私的同时，不影响模型的正常使用。

# 第五章 总结

**项目成果总结**

LLM-GUARD系统作为一种面向线上大语言模型的多模态数据保护工具，成功地解决了当前大语言模型使用过程中面临的敏感信息泄露问题。通过创新的数据保护模式、大文件处理方式和prompt加密方法，系统在安全性、灵活性和易用性方面表现出色。经过严格的测试，系统在功能、性能和安全性方面均达到了预期目标，能够满足用户在不同场景下的数据保护需求。

**项目意义**

随着大语言模型在各个领域的广泛应用，数据安全问题日益凸显。LLM-GUARD系统的开发不仅为用户提供了一种有效的数据保护解决方案，还推动了大语言模型在安全领域的应用和发展。该系统具有重要的现实意义和广阔的应用前景，尤其适用于保密机构、企业、高校核心部门等对数据安全要求较高的场景。

**未来展望**

尽管LLM-GUARD系统在当前版本中已经取得了显著的成果，但仍有一些改进空间。未来，我们计划在以下几个方面进行进一步的研究和开发：

1. **性能优化**：进一步优化系统的响应时间和资源消耗，提高系统在处理大规模数据时的效率。
2. **策略优化**：丰富策略生成模块的功能，提供更多定制化的策略模板，以满足不同用户的需求。
3. **多语言支持**：扩展系统的语言支持范围，使其能够处理更多语言的文本和prompt信息。
4. **用户反馈机制**：建立更完善的用户反馈机制，根据用户的实际使用情况进行持续优化和改进。

总之，LLM-GUARD系统作为一款创新的数据保护工具，有望在大语言模型的安全应用领域发挥重要作用，并为未来的研究和开发提供有益的参考。

# 参考文献

[1]李建华.网络空间威胁情报感知、共享与分析技术综述[J] .网络与信息安全学报，2016, Vol. 2(2): 16-29. （样例，参考国标GB/T7714-2015）

[2] OWASP. OWASP Top 10 for LLM [EB/OL]. (2023-08-15) [2024-03-25]. <https://owasp.org/www-project-top-10-for-large-language-model-applications/>.

[3] OpenAI. ChatGPT: Optimizing Language Models for Dialogue [EB/OL]. (2023-03-14) [2024-03-25]. <https://openai.com/blog/chatgpt/>.

[4] Samsung. Samsung Employees Accidentally Leaked Sensitive Data to ChatGPT [EB/OL]. (2023-04-20) [2024-03-25]. <https://www.samsung.com/global/news/samsung-employees-accidentally-leaked-sensitive-data-to-chatgpt/>.

[5] Gao, Y., et al. PPTX2MD: A Tool for Converting PPT to Markdown [EB/OL]. (2022-07-15) [2024-03-25]. <https://github.com/ygao/pptx2md>.

[6] PDFMiner. PDFMiner: A Tool for Extracting Text from PDF Documents [EB/OL]. (2021-05-10) [2024-03-25]. <https://github.com/pdfminer/pdfminer.six>.

[7] Markitdown. Markitdown: A Tool for Converting Office Documents to Markdown [EB/OL]. (2022-03-20) [2024-03-25]. <https://github.com/markitdown/markitdown>.

[8] ChromaDB. ChromaDB: An Efficient Vector Database [EB/OL]. (2023-01-15) [2024-03-25]. <https://github.com/chroma-core/chroma>.

[9] Ollama. Ollama: A Lightweight LLM Engine [EB/OL]. (2023-04-20) [2024-03-25]. <https://ollama.com/>.