南京农业大学课程考核情况表

2024—2025学年第一学期

学 院：信息管理学院

课程名称：数据标注方法与实践 学分： 1

学号姓名：19122120侯兆晗

考核方式： 期末报告

考核内容： 数据标注

评 语：

成绩：

教师签名：

时间：

**数据标注方法与实践课程总结**

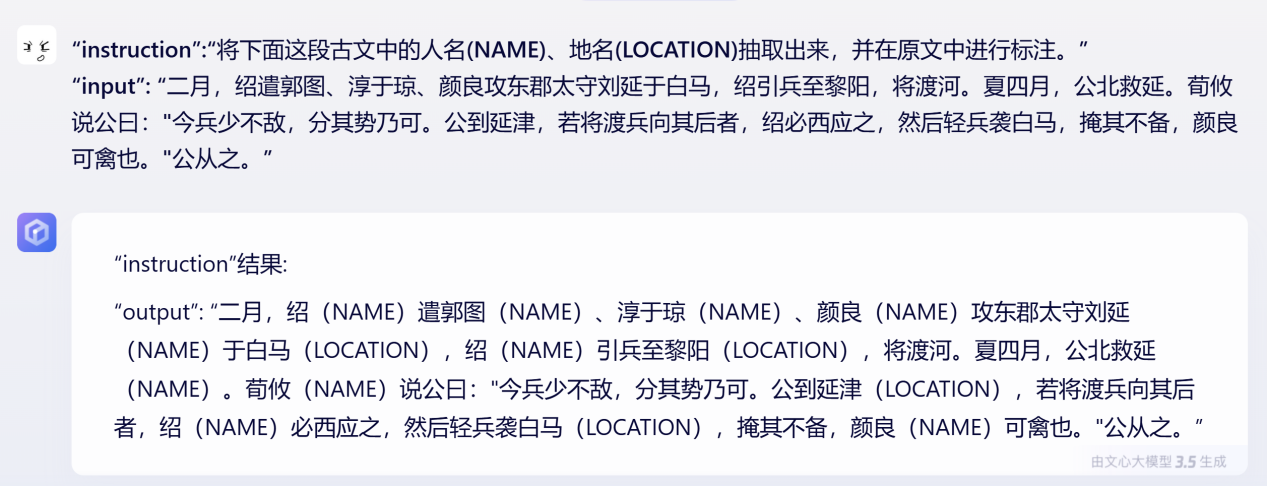
**一、引言**

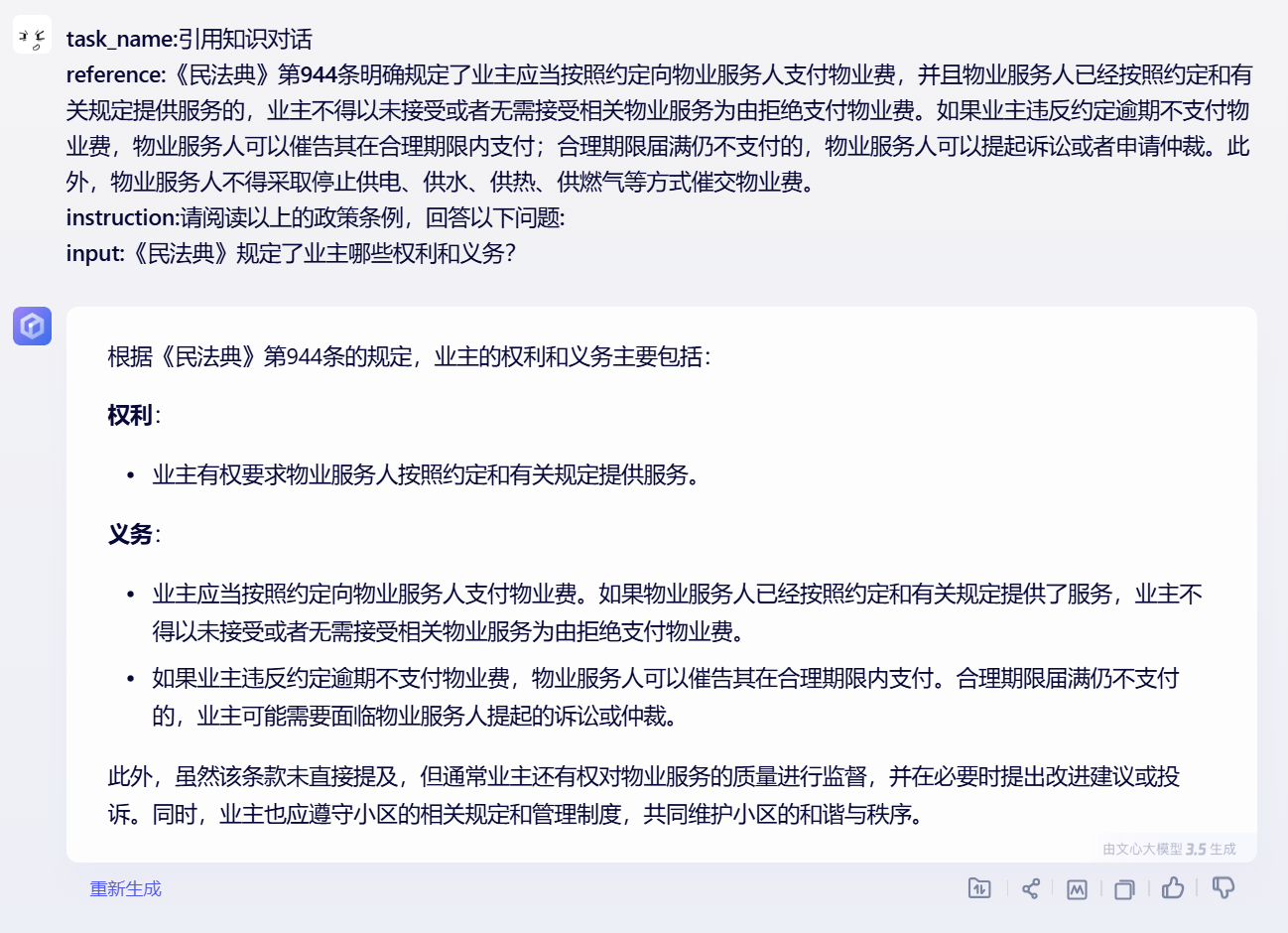
在大三上学期，我学习了《数据标注方法与实践》。这门课程的内容形式新颖，涵盖了数据标注的基本概念、方法、工具和实际应用。王东波和刘浏两位老师通过大语言模型发展前沿、生动的案例和详细的讲解，让我对数据标注有了更加直观的认识。

**二、对课程的认识**

数据标注方法与实践是一门专注于大语言模型领域技术和应用的课程，这门课可以帮我深入理解数据标注的基本概念、方法、工具及其在机器学习中有监督学习的实际应用。这门课可以帮我对下学期即将选修的中文信息处理课程有更好的了解。数据标注在构建高质量数据集、训练和优化模型中的作用愈加重要，我院王东波教授团队通过数据标注的方法训练出的模型，例如荀子大语言模型、中华物产大语言模型受到业内外人士好评。

课程的前三次课刘浏老师和王东波老师为我们讲解了大模型发展技术前沿、大模型的指令跟随、对齐指令设计原则以及思维链与外部知识使用等知识。在这门课上我首先了解了自然语言与程序语言的概念辨析，自然语言通常是指一种自然地随文化演化的语言。是人类交流和思维的主要工具，程序语言是用来定义计算机指令执行流程的形式化语言，程序语言的发展经历了机器语言、汇编语言和高级语言三个阶段。通过数据标注可以实现大语言模型的指令跟随能力使得人类用自然语言控制智能。





上图为利用百度文心一言进行的提示编写。人类通过将自己需要的处理任务转化为提示文本，交由大语言模型进行处理获得反馈。通过总结和了解，大语言训练步骤为因果语言模型预训练、多任务指令微调、人类反馈的强化学习。大模型的数据标注是一个技术含量很高的工作，绝非机械劳动。

当遇到复杂困难问题时，可以将其分解为一些中间问题然后逐步解决，最后得到最终结果，也称为“自顶向下，逐步求精”(top down stepwise refinement)。思维链(Chain-of-Thought)就是参考人类解决问题的方法，从输入问题开始的一系列自然语言形式的推理过程，直到得到最后输出结论。和之前的提示词模式(问题，答案)不同，思维链提示词模式即(输入问题、思维链、输出结论)。模型在输出部分中先进行了一段演绎推理，然后才输出我们需要的结果。

总结上课内容可以学到关于标引的知识。标引的主要流程包括：（1）对文献进行主题分析，查明其中所论的有情报价值的内容，将这些内容概括为若干主题概念；（2）根据检索系统使用者的需要，决定对该文献采取何种标引方式，即对主题概念作出取舍和处理的决定；（3）根据所采用的分类表或词表，按照决定采取的标引方式，将主题概念转换成检索标识——分类号或检索词；（4）按照所给予文献的检索标识，将文献的书目记录组织成检索系统。标引主要分为抽词标引和赋词标引，抽词标引指自动抽词标引是由计算机根据词频分布等特点，从文献的标题或正文中提取能表达该文献主题的关键词或词组，用以标识文献主题的过程；赋词标引指将文献中表达主题内容的有用词转换成体系的受控词的过程。标引任务有四点：（1）为文本生成简单的二次文献，可以是对原文的简述，也可以是对原文中发生事件的梳理。（2）整理思维链，根据生成的二次文献和文献来源共同推理出文本的背景。（3）根据演绎所得文本，生成文章的关键词(一般5个)。关键词需来源于文本。（4）根据文本内容及推演内容，生成文章的主题词(一般5个) 。主题词无需来自文本，但需要有概括性。

**三、学习困难与自我剖析**

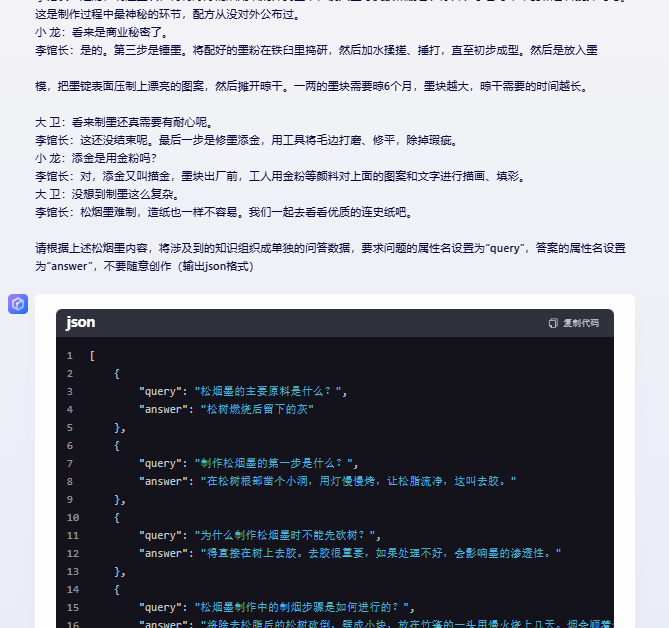
1.技术小白的转变

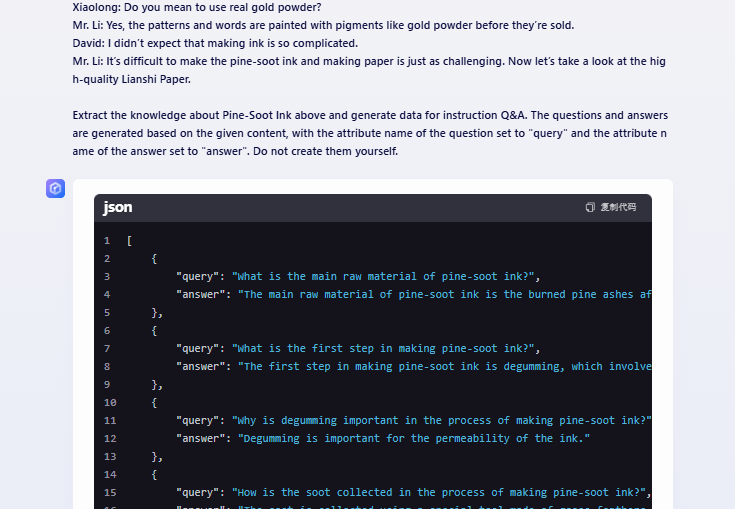
在课程开始之初，数据标注对我来讲是一个全新的领域，我对数据标注知识几乎一无所知，是个不折不扣的技术小白。面对各种标注工具、方法和流程，我感到既新奇又有些手足无措。

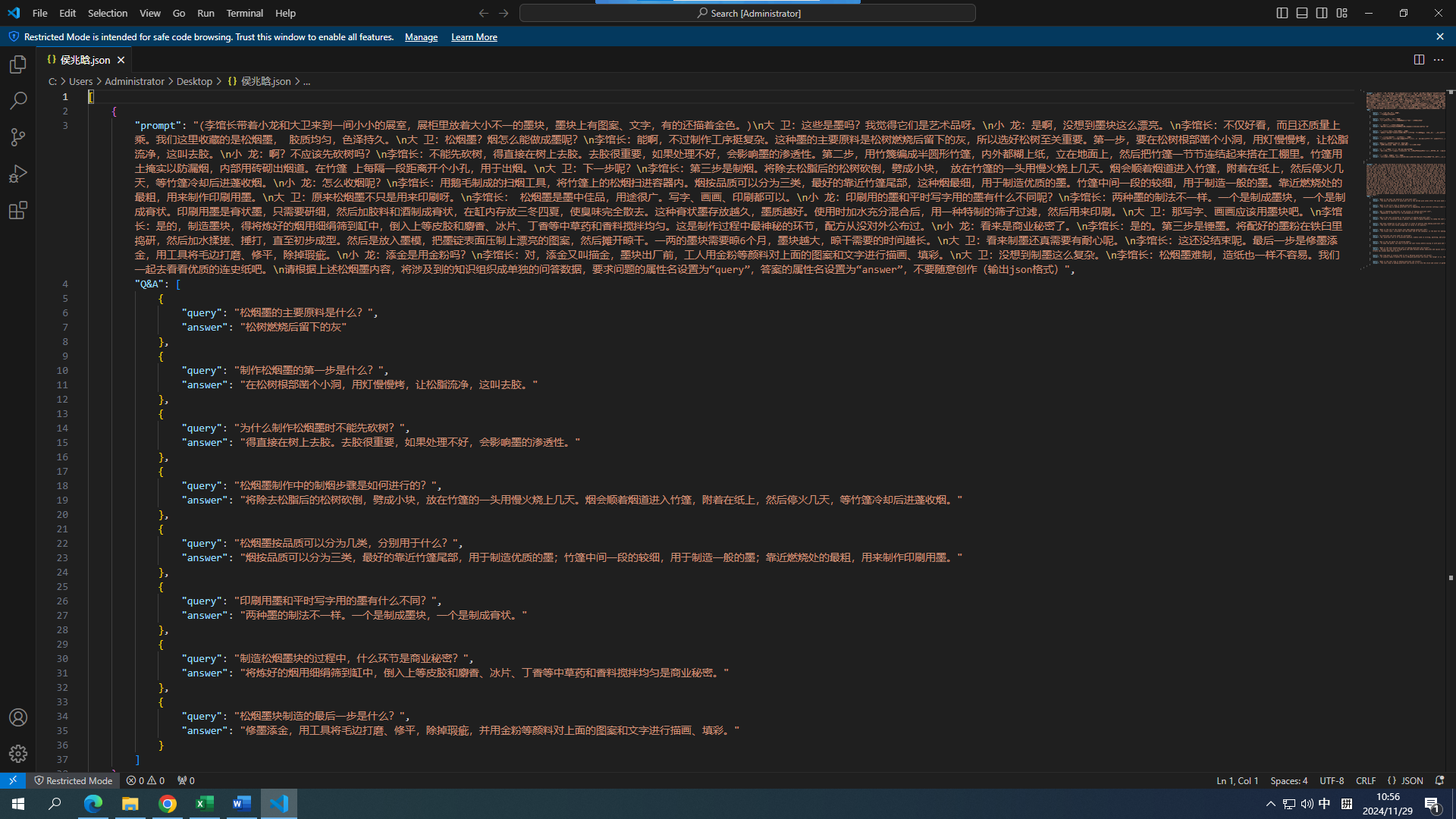
在我初次接触数据标注课程时，对这个领域充满了好奇与迷茫。面对海量的数据、复杂的标注规则和严格的准确率要求，我感到既兴奋又有些手足无措。那时，我意识到，要想在这个领域有所作为，必须付出努力，不断学习与实践。

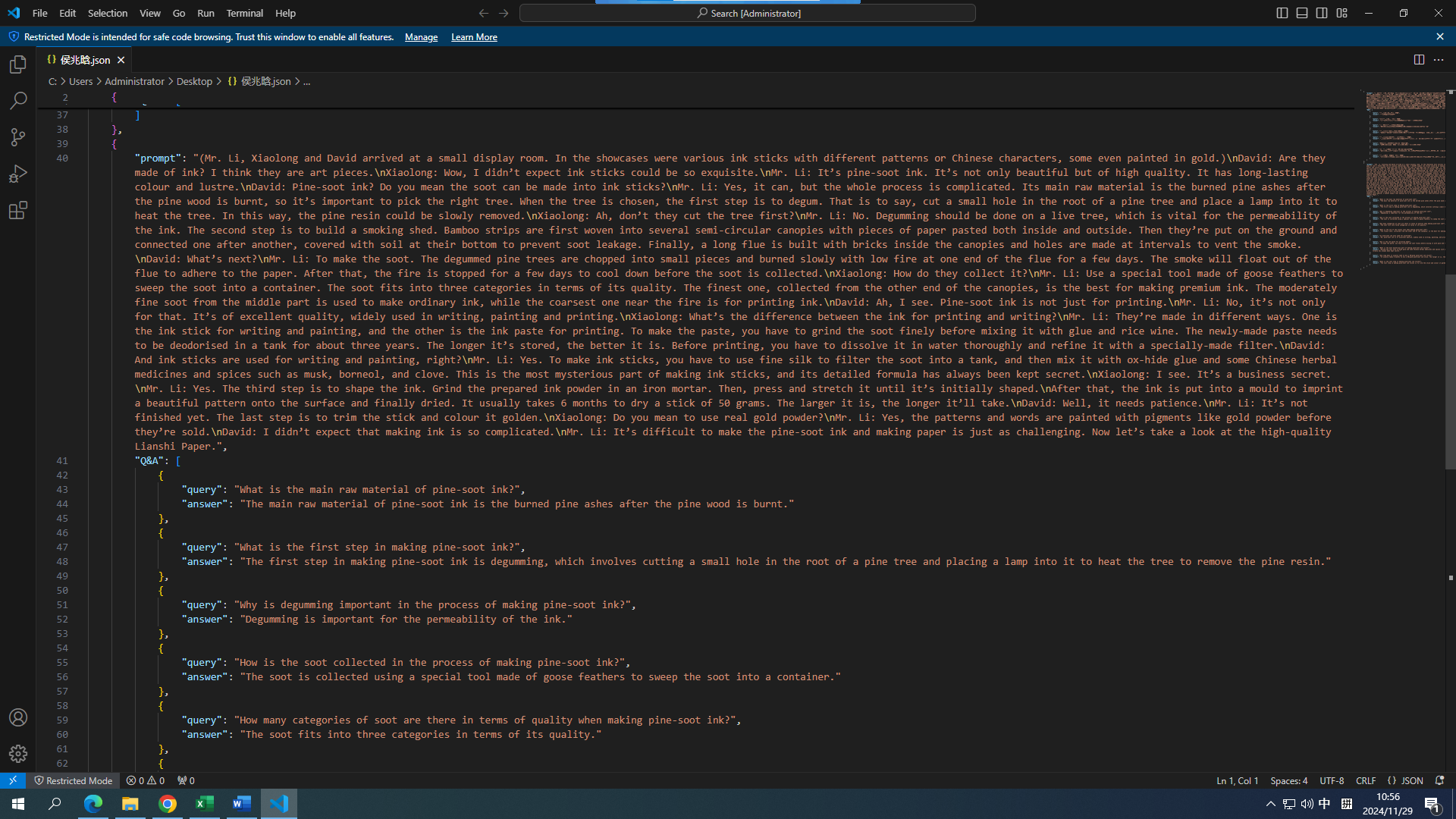
起初，课程从最基础的标注任务开始，如文本中的关键词提取等。这些看似简单的任务，实际上却考验着我的耐心、细心和对标注规则的准确理解。我逐渐学会了如何高效地阅读标注规范，如何快速识别并标注数据中的关键信息，以及如何确保标注的一致性和准确性。

随着刘浏老师和王东波老师讲解的不断深入，使我够深入理解文本背后的含义和情感色彩。为了提升标注质量，我开始主动学习相关的自然语言处理知识，如分词、词性标注、句法分析等，这些知识不仅拓宽了我的视野，也为我解决标注中的难题提供了有力的支持。









2.遇到的困难

当我第一次踏入数据标注课程的大门时，内心对于对这门课感到了一丝忐忑。尽管第一次刘浏老师上课时我尚能维持良好的学习步伐，但每当触及技术，特别是那些与数据标注紧密相关的工具和方法时，我的自信心便会有所动摇。数据标注课程充斥着大量的数据生成、标注及人工校对的后续处理工作，而首要难关便是如何以高效的方式完成数据的人工分析和校对。

通过课后坚持不懈的实战演练，我逐渐熟悉了基础的标注工具Label Studio平台，并深入了解了标注的规范标准。起初，我面临的挑战主要集中在理解标注全部的流程和大语言模型如何使用这些标注好的数据。面对技术上的重重障碍，我通过积极向研究生助教请教以及与同学交流探讨，逐步积累了丰富的经验，从而强化了自己的技术能力。随着标注水平的日益精进，我也学会了如何规避常见的标注失误，特别是在确保标注质量的同时，也兼顾到了效率的提升。这段经历使我深刻认识到王东波教授一直强调的话：数据标注是一门实践性很强的课程。

**四、学习内容与技巧**

1.学习目标

在学习这门课时，为尽快学习、了解和应用数据标注的技术，我制定了相应的目标：（1）理解数据标注的基本概念：能够清晰地定义数据标注，并理解其在数据处理和分析中的重要性。（2）掌握数据标注的基本工具和技术，并掌握标注数据的格式转换和存储方法。（3）学习如何制定有效的标注策略，以提高标注的准确性和一致性，从而提高标注效率。

2.多方面了解新技术

为了快速入门并跟上课程进度，我采取了多方面了解新技术的策略。除了认真听讲、阅读上课ppt外，我还积极浏览相关论坛、博客，观看教学视频，这些努力让我对数据标注有了更全面、更深入的理解。

在本课程的实验环节中，我专注于掌握数据标注的时间步骤的练习。通过参与“大语言模型术语”与“非物质文化遗产（非遗）术语”等实际标注实践，我不仅深入掌握了Label Studio平台的标注技术，还精通了质量把控与数据导出json格式的方法。我全面学习使用Label Studio平台，熟练掌握了从创建标注项目、数据导入到标签设定等一系列基本操作。该平台功能全面且操作简便直观，为我的标注任务提供了高效且强大的支持。

3.充分实践

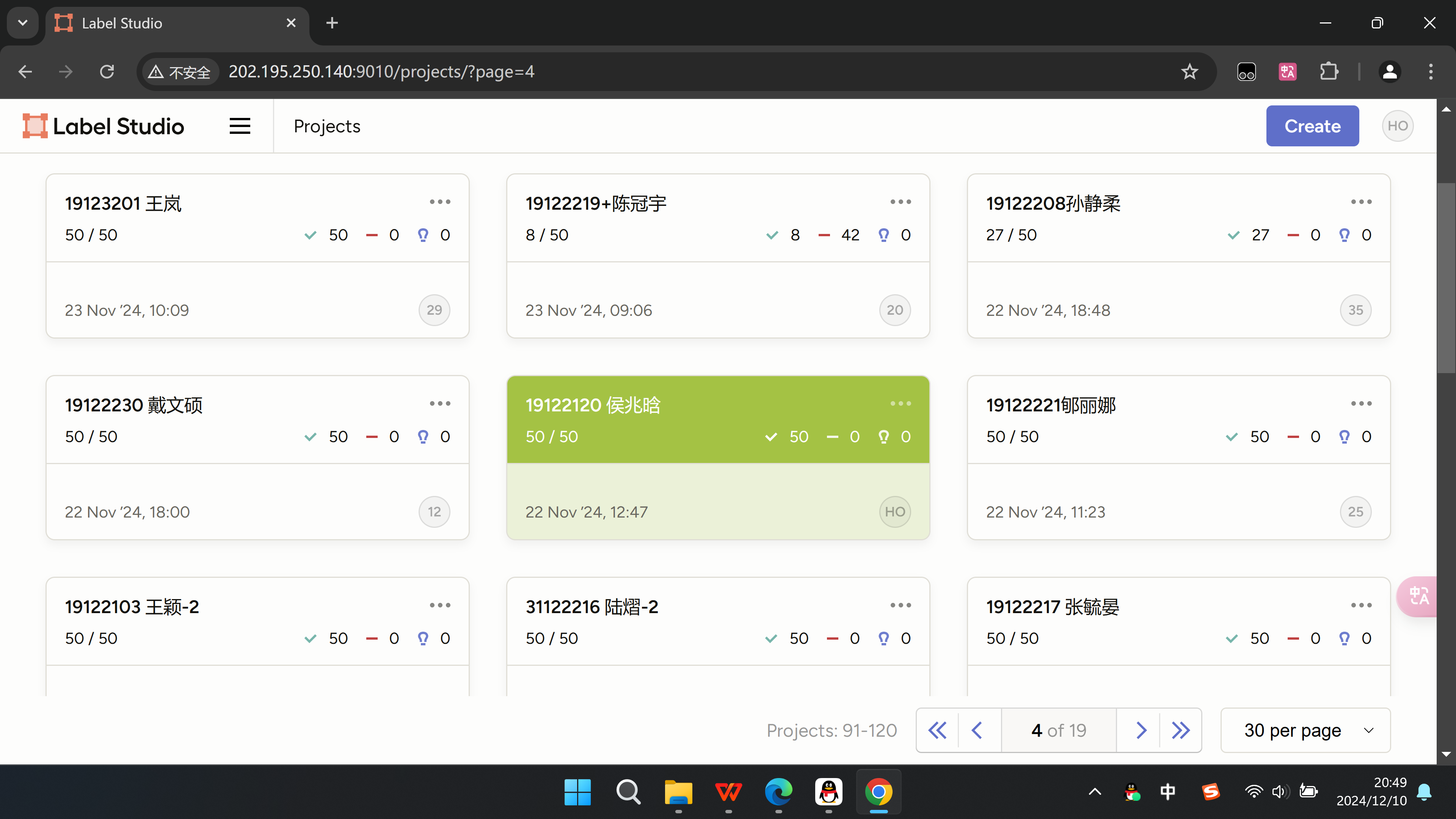
王东波教授强调这门课是实践动手能力要求比较高的课。理论知识固然重要，但实践才是检验真理的唯一标准。因此，在学习过程中，我始终注重将所学知识应用于实际项目中。经过一系列重复的标注实践，我掌握了将理论知识付诸实际操作的技巧。对此我总结出Label Studio平台实践方法步骤：

（1）注册登录

我使用电子邮箱注册并登录了Label Studio平台，并根据要求配置了相关的实验设置。通过这一过程，我熟悉了一些基本操作界面和功能，为后续的标注工作做好了前期准备。

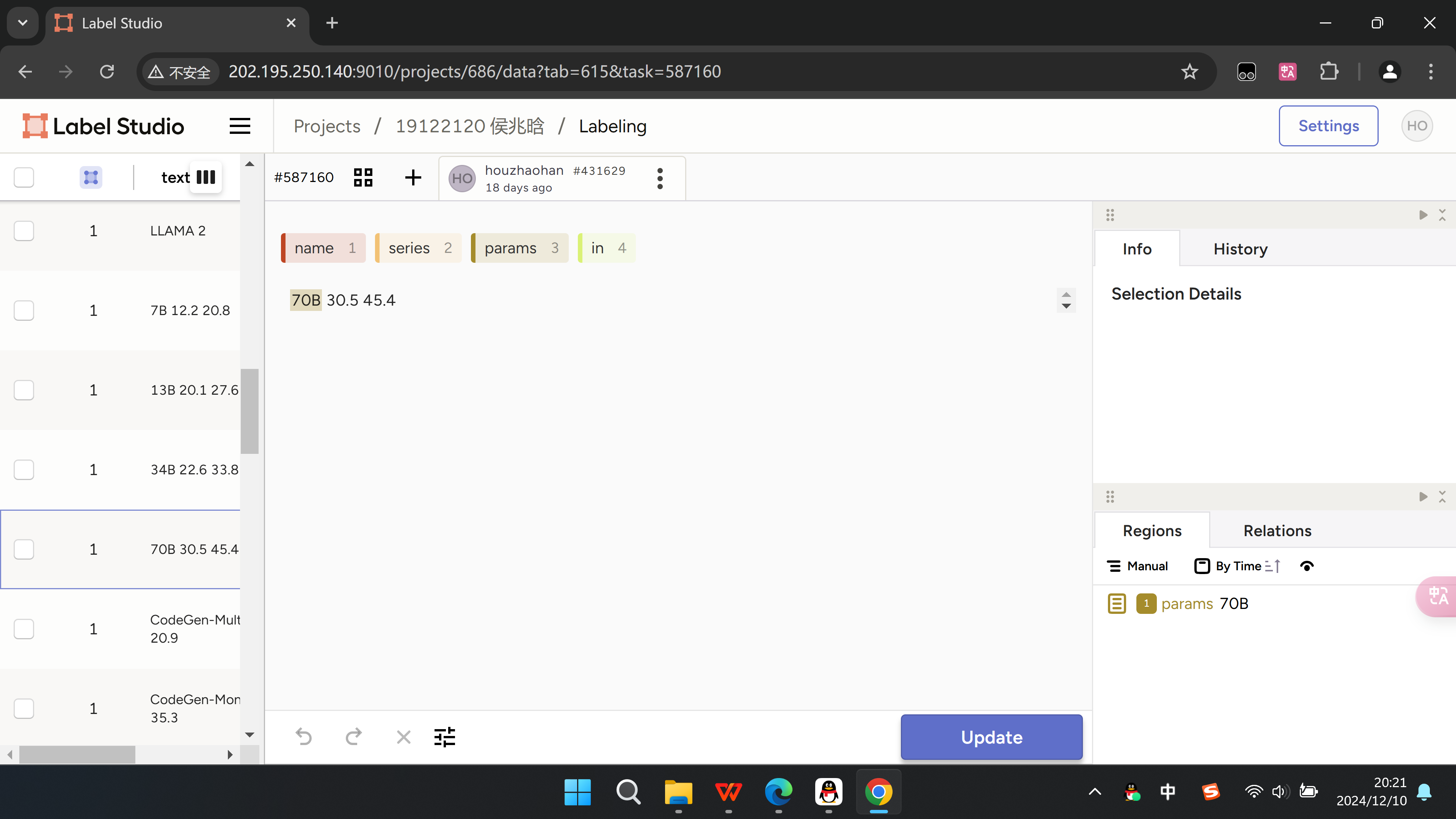
（2）项目创建

根据说明文件指导，我创建了自己的项目，并把txt文件导入进Label Studio，并创建了项目名称。



（3）设置实体标签

选择natural language processing，点击name entity recognition，设置实体标签，本次实体标注使用四大类实体，模型名称</name>、模型系列</series>、参数量</params>、开发机构</in>。

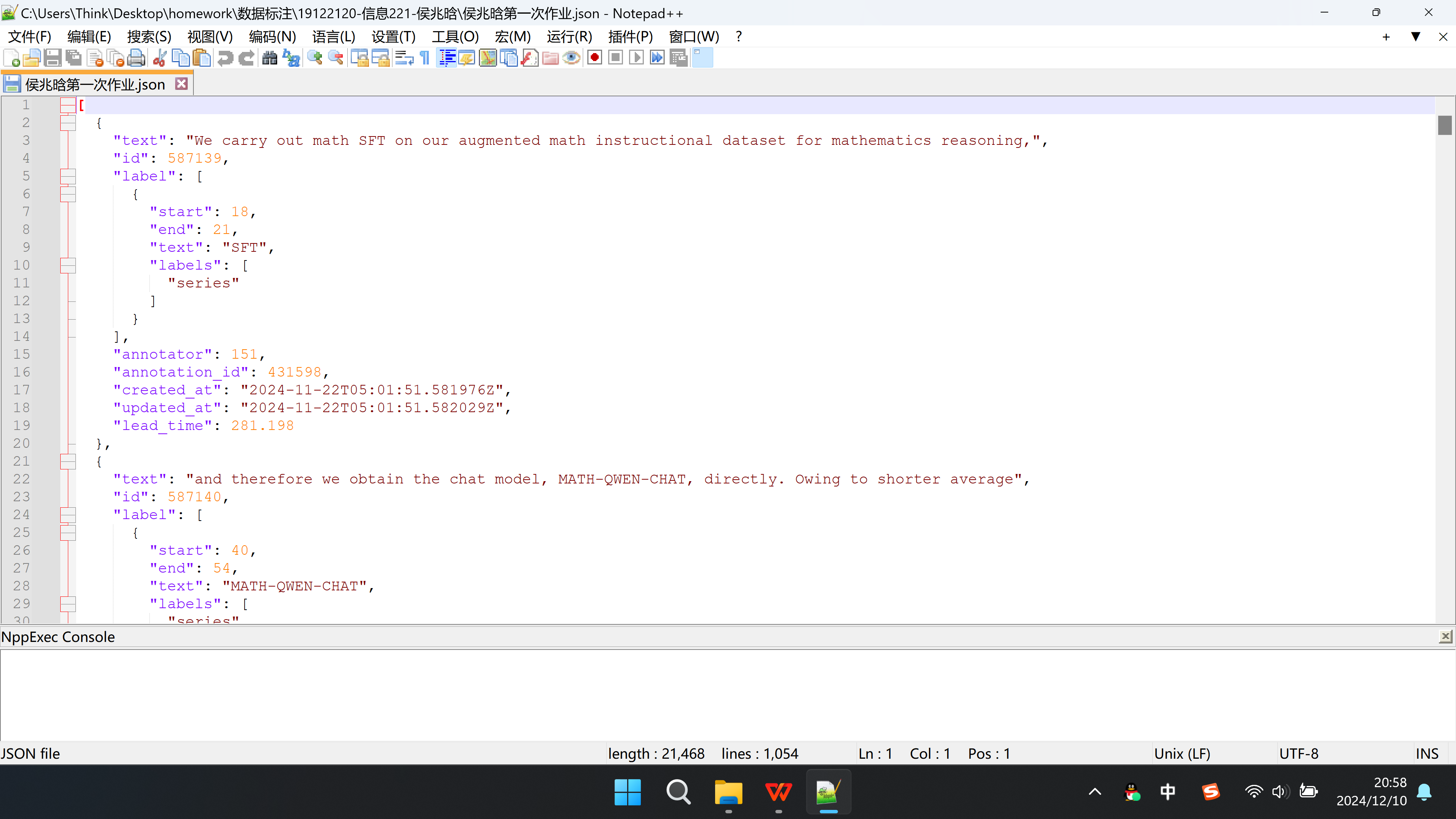


（4）进行标注

根据阅读“大模型术语标注规范.pdf”，对内容进行判断后标注，标注过程中，我严格按照规范，确保标注内容正确。

（5）结果导出

对一个项目标注完成后，进入项目页面，按照json-mini格式进行导出，即可下载导出json文件。



对于以上标注过程来讲，在进行标注时，我的工作不再局限于机械式的任务执行，而是融入了每一步的深思熟虑，剖析其中的模式并探索可能的改进途径。我领悟到，确保标注的精准固然重要，但维持标注的一致性和提升效率同样不可忽视，尤其是在后续的数据分析阶段，这两个要素直接关系到数据的可用度和可信度。

在标注实践中，我还学会了借助工具来提升工作效率。以Label Studio平台为例，我不仅着眼于单次标注的结果，更通过合理配置实体标签、制定标注规范及优化工作流程，显著提高了标注的质与量。这些经历使我逐步精通了数据标注的核心技巧，为将其有效应用于实际项目奠定了坚实基础。

五、实践总结反思

在实践环节，我深刻体会到了数据标注的复杂性和挑战性。从最初的大语言模型交互到后来的文本标注等，每个项目都让我受益匪浅。通过实践，我不仅掌握了各种标注工具的使用方法，还学会了如何根据项目需求制定合理的与大语言模型交互的方法。同时，我也深刻认识到了数据质量对有监督学习模型性能的重要性，因此在标注过程中始终保持着高度的责任心和严谨的态度。

最后，我不仅在这门课上数据标注方法，还培养了利用大语言模型解决实际问题的思维方式和技巧。虽然这门课只有6周，但通过不断的学习和思考，我也感受到了自己在数据标注领域的成长和进步。未来，我希望能够继续深入学习数据标注的相关知识，并将其应用于更广泛的领域，为人工智能的发展贡献自己的力量。