

"编译技术"课程大语言模型使用指导和约束报告

引言

随着人工智能的时代浪潮出现，与大语言模型（Large Language Model，简称LLM）的飞速发展与生成质量的提高，同学们也许已经在编写代码，撰写报告中或多或少养成了依赖大模型的趋势，又或许还并未尝试过使用大模型辅助完成各项任务。无论如何，为顺应时代的发展，本报告旨在指导学生在编译技术课程中合理、高效地利用大模型来支持理论学习和编译器开发，同时明确使用的界限，以确保学术诚信和培养学生个人思维能力。

大语言模型使用指导

1. 推荐使用的大模型

随着国内大模型技术的迅猛发展和激烈竞争，众多模型开始提供对多模态数据和长文本处理的先进支持。为了帮助同学们在学习过程中更有效地解决具体问题，如针对编程题目截图或编译文档的分析和提问，课程组精心推荐以下三款国产免费大模型，它们将极大地助力同学们的课程学习。

1. [Kimi](#)：Kimi模型是月之暗面公司推出的大模型服务，也是作者经常使用的国产大模型之一，其率先支持多模态文件读入与长文本总结能力，使用效果极佳，总部就在北航门口。
2. [通义千问](#)：通义千问是阿里云推出的大模型服务，如今其上下文理解能力和支持上下文长度已领先Kimi(2024.06更新)，成为了目前中文大模型的领跑者之一，使用效果同样极佳。
3. [智谱清言](#)：智谱清言是智谱AI推出的ChatGlm系列大模型服务，推荐使用Glm-4版本，其开放平台支持构建专属知识库，适合用于构建特定领域的问答系统，或许可辅助期末复习。

如有条件，课程组同样推荐使用GPT-4，GPT-4o，Claude3.5，Llama3.1，Gemini1.5等国外大模型，课程组认为它们可能具备更强大的自然语言理解与代码生成能力，但请同学们注意科学上网与保护个人隐私安全，自行探索使用方法。

上述大模型均经过测试，可以一次性较完整的实现 [样例编译器](#) 的词法与语法解析工作。

2. 学习和理解理论知识

课程组鼓励同学们将大模型作为课堂之外的辅助学习工具，帮助学生理解编译原理中的抽象概念和复杂理论，包括语法分析、语义分析、中间表示、代码优化等。同学们可以构造合适的提示词向LLM提出关于理论知识的具体问题，例如：

“LR(0)分析器有什么问题，可以如何改进”：

🌟 Claude-3.5-Sonnet Poe ...

LR(0)分析器虽然是一种强大的语法分析技术，但它确实存在一些限制和问题。让我们来探讨这些问题以及可能的改进方法：

LR(0)分析器的主要问题：

1. 移进-规约冲突：
当分析器无法决定是应该移进还是规约时，会发生这种冲突。
2. 规约-规约冲突：
当分析器面临多个可能的规约选择时，会发生这种冲突。
3. 有限的表达能力：
LR(0)分析器无法处理某些类型的文法，特别是那些需要向前看符号来做出决策的文法。
4. 大型分析表：
对于复杂的文法，LR(0)分析表可能变得非常大，这会增加内存使用和查找时间。

改进方法：

1. SLR(1)分析器：
 - 使用 FOLLOW 集来解决一些冲突。
 - 可以处理更多的文法，同时保持较小的分析表。
2. LALR(1)分析器：
 - 结合了 LR(0) 项目集和向前看符号。
 - 比 SLR(1) 更强大，可以处理更多的文法。
 - 分析表大小与 LR(0) 相当，但能力接近 LR(1)。
3. LR(1)分析器：
 - 使用一个向前看符号来解决冲突。
 - 可以处理几乎所有的确定性上下文无关文法。
 - 缺点是分析表可能变得非常大。
4. GLR分析器（广义LR分析器）：
 - 可以处理所有上下文无关文法，包括二义性文法。
 - 在遇到冲突时，会并行探索所有可能的解析路径。
 - 适用于自然语言处理等需要处理二义性的领域。

在解答理论题目方面，经测试简单题基本可以一次给出正确的答案与解释，而复杂题目诸如构造分析表等需要多次引导才能给出接近正确的解答，建议以参考答案和同学间交流为主。例如：

"请填空"（正确解答）：

7. 对文法 $G[T]: T ::= T-T \mid T/T \mid (T) \mid i$ ，规范句型 $T-T/i$ 的句柄为_____和_____, 由此判断该文法_____（有/无）二义性。
8. 规范归约每次归约的是句型的_____, 算符优先分析法每次归约的是当前句型的_____。
9. 活动记录中 Display 区存放的是_____。
10. 文法 $G = (V_n, V_t, P, Z)$ ，其中 V_t 代表_____。

以下是填空的答案：

7. 对文法 $G[T]: T ::= T-T \mid T/T \mid (T) \mid i$, 规范句型 T/Ti 的句柄为 **T/T** 和 **i** , 由此判断该文法 **有** (有/无) 二义性。
8. 规范归约每次归约的是句型的 **最左素短语**, 算符优先分析法每次归约的是当前句型的 **最左素短语**。
9. 活动记录中 Display 区存放的是 **指向外围活动记录的静态链指针**。
10. 文法 $G=(V_n, V_t, P, Z)$, 其中 V_t 代表 **终结符集合**。

"请解答这道题目" (反面教材) :

3. 给定文法

$$\begin{aligned} N &\rightarrow MN' & P &\rightarrow QP' \\ N' &\rightarrow iN \mid \varepsilon & P' &\rightarrow +P' \mid \varepsilon \\ M &\rightarrow PM' & Q &\rightarrow (N) \mid a \mid b \mid \wedge \\ M' &\rightarrow M \mid \varepsilon \end{aligned}$$

- (1) 求各非终结符的 FIRST 集和 FOLLOW 集 (3 分)
- (2) 请说明 LL(1) 的充分必要条件, 并判断上述文法是否为 LL(1) 文法 (3 分)
- (3) 构造该文法的分析表。请直接填写下页表格。(10 分)

总结 FOLLOW 集:

```
scheme 复制
FOLLOW(N) = {' ', '$'}
FOLLOW(N') = {' ', '$'}
FOLLOW(M) = {'i', ')', '$'}
FOLLOW(M') = {'i', ')', '$'}
FOLLOW(P) = {'i', ')', '$'}
FOLLOW(P') = {'i', ')', '$'}
FOLLOW(Q) = {'i', ')', '$'}
```

请注意, 即使如今以 GPT-4o 为首的大模型对计算机领域相关知识理解很深, 但也应注意分辨真伪, 同学们应当与编译教材, 教师课堂授课内容, 授课 PPT 相结合批判性思考, 遇到矛盾的地方应以课堂所学为基准。

3. 编程和项目开发

经过面向对象课程的学习, 想必同学们对于类和对象有了更深刻的理解, 本课程的核心任务——编译器的开发需要用到众多编程知识。课程组同样鼓励同学们使用大模型辅助同学们的编译器开发工作, 例如以下:

1. **语法和编程问题**：在开发编译器过程中，鼓励同学们咨询大模型有关编程语言的语法问题与报错信息等，提高代码编写与Debug的效率。
2. **编程思路和架构优化**：同学们可以尝试使用大模型帮助优化自己的编译器架构，或使用大模型学习参考经典的编译器设计架构与子架构。
3. **★代码编写辅助工具★**：课程组设计实验旨在考察编译器设计思想而非为难于编程实现，因此同样鼓励同学们探索AI编程助手工具辅助开发，比较流行的Copilot, Tabnine, Kite等需要翻墙且付费，同学们可自行学习使用，较完整的编程助手推荐参见[链接](#)。在这里课程组仍推荐两款国产免费编程助手，[CodeGeeX](#)(配置参考[链接](#))与[通义灵码](#)(配置参考[链接](#))，使用得当将大幅提升同学们的开发效率，祝愿大家在开发过程中一帆风顺。

再次特别强调，课程组以激发同学们的求知欲探索欲为主要目标，同学们应主动尝试解决问题，锻炼自己的逻辑思维，并将大模型的建议和反馈作为参考，而非完全依赖。要懂得“**如何更好使用工具**”，而非“**被工具操纵**”。

使用约束

在本课程所要求的一切文字工作上，课程组有明确的**禁止使用大模型约束**。

本课程的写作工作包括阶段设计文档，总设计文档，优化文档以及课程总结报告。写作任务是评估同学对课程内容理解和思考的关键组成部分，必须由同学独立完成。同学们在提交的任何文字材料中都应保持原创性，若存在少量使用大模型生成的内容，**必须要**清楚标注在文档内以供课程组评估。

惩罚措施

为切实落实该指导与约束报告，课程组对此具有一套完整的惩罚措施。代码方面，课程组具有完备的代码查重机制；文档方面，课程组将使用软件审查+人工检查的双审查模式检查同学们写的每一份文档，若发现存在使用大模型代写的嫌疑，课程组会立刻联系相应同学，并查清原委。

丑话说在先，违反者将被扣除部分实验分数作为处罚，倘如多次出现屡教不改情况，则直接**取消实验成绩**，望各位同学谨遵规范！

作者：杨博文助教

发布时间：2024.09.15