

大规模语言模型增强的中文篇章多维度阅读体验量化研究

孙嘉黛, 汤思怡, 王诗可, 于东*, 刘鹏远

北京语言大学/ 信息科学学院

北京市海淀区学院路15号, 100083

{suesunegg,tangsiyi0805,shikewang98}@gmail.com

yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

摘要

现有的文本分级阅读研究往往从文本可读性的角度出发, 以离散的文本难度等级的形式为读者推荐阅读书目。目前, 仍缺少一种研究读者在阅读过程中产生的多方面、深层次阅读体验的体系结构。对此, 我们调研了读者在阅读中文篇章过程中产生的不同阅读体验, 提出了中文篇章多维度阅读体验的量化体系。我们将阅读过程中呈现的连续性的阅读体验归纳为多种类别, 并在此基础上构建了中文篇章多维度阅读体验数据集。同时, 我们探究了以大规模语言模型为基础的ChatGPT对阅读体验的量化能力, 发现其虽具备强大的信息抽取和语义理解能力, 在阅读体验的量化上却表现不佳。但我们发现大规模语言模型所蕴含的能力能够以知识蒸馏的方式协助深层属性的量化, 基于此, 我们实现了大规模语言模型增强的中文篇章多维阅读体验量化模型。模型在各维度阅读体验上的平均F1值达到0.72, 高于ChatGPT的Fewshot结果0.48。

关键词: 阅读体验; 大规模语言模型; 知识融入

Quantitative Research on Multi-dimensional Reading Experience of Chinese Texts Enhanced by Large Language Model

Jiada Sun, Siyi Tang, Shike Wang, Dong Yu*, Pengyuan Liu

Beijing Language and Culture University / Faculty of Computer Science

15 Xueyuan Road, Haidian District, Beijing, 100083

{suesunegg,tangsiyi0805,shikewang98}@gmail.com

yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

Abstract

Existing studies on graded reading often recommend reading materials to readers in the form of discrete difficulty levels from the perspective of text readability. However, there is still a lack of a systematic framework that can capture the multi-faceted and deep reading experiences of readers during the reading process. To address this issue, we investigate the different reading experiences of readers during the process of reading Chinese texts and propose a quantitative framework for multidimensional reading experiences in Chinese discourse. We group the continuous reading experiences that emerged during the reading process into multiple categories and constructe

*为通讯作者

基金项目: 教育部人文社会科学基金项目(19YJCZH230); 中央高校基本科研业务费(北京语言大学梧桐创新平台, 21PT04)

©2023 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

a dataset of multidimensional reading experiences in Chinese discourse. Additionally, we explore the ability of ChatGPT, based on Large Language Models, to quantify reading experiences and find that while it has strong information extraction and semantic understanding capabilities, it performs poorly in quantifying reading experiences. However, we find that the capabilities inherent in Large Language Models can assist in the quantification of deep attributes through knowledge distillation. Based on this, we implement an enhanced model for quantifying multidimensional reading experiences in Chinese discourse using a Large Language Model. The model achieves an average F1 score of 0.72 on various dimensions of reading experience, which is higher than ChatGPT's Fewshot result of 0.48.

Keywords: Reading Experience , Large Language Models , Knowledge Intergration

1 引言

在全民阅读氛围浓厚的当下，阅读推广活动正在广泛开展，中小學生等青少年兒童在被推广人群中的比重最高。由此，为青少年推荐适合他们阅读的优秀读物是专家和学者一直关注的话题。为了适应匹配不同阅读水平的青少年读者，分级阅读也成为近年来研究的重点(Rao et al., 2021)。目前，分级阅读通常只考虑文本难度，即文本可读性(Dale and Chall, 1948)。但阅读也需要深度和广度，在保障难度适配的同时增加分级阅读文本的多样性、提升读者在阅读过程中的体验感也是阅读研究的新兴方向(池春红, 2022)。何应艳(2016)认为阅读体验即为读者在理解文字的情况下，通过阅读产生的深层次、多元的感受。在此基础上，我们总结了阅读体验的两个特征，一是深层次，二是多维度。因此，阅读体验的概念可以扩展为读者在阅读过程中对文本的不同维度产生的多元感受。与文本难度相比，阅读体验结合了文本客观因素、读者心理因素两方面，它更加重视阅读过程中读者的主观感受。

文本可读性研究往往从字数、词数、句法结构、篇章结构等浅层语言特征对文本进行分析，忽视了文本的一些深层语义内涵，也忽视了读者与文本的互动性。图1展示了两个来自同年级语文课文的不同文段，两个文段在文字难度上基本相同，即在可读性上的等级相同，但两者给读者带来的阅读体验却截然不同。文段1通过优美的语言表达，给读者较高的文采体验；而文段2用词平实，却讲述了一个具有哲理性的故事，更加引人深思。融合阅读体验进行文本分析、分级阅读、读物推荐等工作，可以从深层与浅层多个角度提取出文本特征，使读者能更快速、准确地感受文章情感，欣赏文章的优美文字，领悟文章的内涵和主旨(杨春and 张秋月, 2022)。

文段1: 道两旁的法国梧桐树，掉下了一张张金黄金黄的叶子。这一张张闪着雨珠的叶子，一掉下来，便被紧紧地粘在湿漉漉的水泥道上。我走在院墙外的水泥道上。水泥道像铺上了一块彩色的地毯，这是一块印着落叶图案的，闪闪发光的地毯，从脚下一直铺到很远很远的地方，一直到路的尽头……每一张法国梧桐树的落叶，都像一个金色的小巴掌，熨帖地、平展地粘在水泥道上。它们排列得并不规则，相反，很凌乱。然而，这更增添了水泥道的美。我一步一步小心地走着，我一张一张仔细地数着。我穿着一双棕红色的小雨靴。你瞧，这多像两只棕红色的小鸟，在秋天里变得金黄的叶丛中，愉快地欢跳着、歌唱着……

文段2: 孙中山想，这样糊里糊涂地背，有什么用呢？于是，他壮着胆子站起来，问：“先生，您刚才让我背的这段书是什么意思？请您给我讲讲吧！”这一问，把正在摇头晃脑高声念书的同学们吓呆了，课堂里霎时变得鸦雀无声。先生拿着戒尺，走到孙中山跟前，厉声问道：“你会背了吗？”“会背了。”孙中山说着，就把那段书一字不漏地背了出来。先生收起戒尺，摆摆手让孙中山坐下，说：“我原想，书中的道理，你们长大了自然会知道的。现在你们既然想听，我就讲讲吧！”先生讲得很详细，大家听得很认真。后来，有个同学问孙中山：“你向先生提出问题，不怕挨打吗？”孙中山笑了笑，说：“学问学问，不懂就要问。为了弄清楚道理，就是挨打也值得。”

图 1. 可读性相同但阅读体验不同的文段

现阶段对阅读活动的探索仅仅着眼于文本可读性的定量研究或基于机器学习、深度学习模型的语义研究，而忽略了其他深层阅读体验的挖掘，基于此形成的分级阅读和文本推荐在广泛性、多样性及可解释性上有所欠缺。对于文本浅层性质的研究大多指向各类可读性公式，许多可读性公式对多种浅层语言特征进行了定量分析(程勇et al., 2020; 刘苗苗et al., 2021)；对于文本深层阅读体验的研究缺失，和文本深层属性相关的研究大多以领域内的分类任务为方向，并

不涉及阅读过程体验感的探索。作为阅读活动的主体，读者的阅读感受能够直观地反映阅读活动中他们的收获，探究读者在阅读过程中基于文本产生的深层次阅读体验能够在分级阅读的基础上融入更多有价值的阅读信息。因此，中文文本多维度阅读体验体系缺失的问题亟待解决。在这样的背景下，我们结合调研分析，将具备连续性的多维阅读体验归纳为特定的类别，并由此构建了中文篇章多维度阅读体验量化体系和数据集。同时，我们发现大规模语言模型出色的信息抽取和语义理解(Bang et al., 2023)能力能够使其表达的自然语言信息能很好地融入模型中以提高模型性能，但在具备多样性的阅读体验量化任务上的表现却不佳。因此，我们提出了以大规模语言模型辅助多维度阅读体验量化的方法，并采用了蒸馏大模型知识的方式来完成这一过程。

本文的贡献主要包括以下三个方面：第一，我们提出了中文篇章多维度阅读体验的量化体系，并基于此构建了中文篇章多维度阅读体验数据集，为衡量中文篇章的阅读体验提供了理论基础，数据集公开在<https://github.com/blcunlp/CMRED>；第二，我们探究了大规模语言模型ChatGPT(Dai et al., 2022)在阅读体验上的能力，发现ChatGPT虽然能以自然语言的形式出色地完成抽取阅读体验相关表达的工作，但在多维度阅读体验数据集上体现的量化能力较差；第三，基于探究的结果，我们提出了大规模语言模型增强的中文篇章多维度阅读体验量化方法，利用了ChatGPT在语义层面的信息抽取能力，以三种不同的方式在模型中融入来自ChatGPT的知识并通过实验验证了方法的有效性。

2 相关工作

文本作为作者向读者传递信息的载体，有着许多不同的属性，在自然语言处理领域讨论的文本属性多集中在文本可读性领域。Dale and Chall (1948)将文本的可读性定义为文本易于理解和阅读的程度和性质，通常被狭义理解为文本的难度。早期的可读性研究尝试将文本词汇层面和句子层面的如词长、常用词比例、句长等浅层特征进行量化，并以此类特征对文本的难度进行回归预测。常用的可读性公式有Reading Ease Score (Flesch, 1948)、Grade Level (Kincaid et al., 1975)、Lexile Framework (Smith and others, 1989)等。然而，可读性公式缺少对语义、篇章层面的深层语言特征的考量(McNamara et al., 2010)，仅以统计学上的相关浅层特征作为构建基础，有效性值得商榷(Schriver, 2000)。此后随着计算机学科的发展，可读性领域也引入了机器学习、深度学习等新方法。此类方法依靠心理语言学、语料库语言学等理论支撑，以计量的方法表示出更多语言特征，通过模型对文本的难度等级进行预测。常用的机器算法有决策树(Decision Tree, DT)、随机森林(Random Forest, RF)、逻辑回归(Logit Regression, LR)、最近邻(K Nearest Neighbor, KNN)、支持向量机(Support Vector Machine, SVM)等(Wu et al., 2018)。深度学习方法则旨在运用深层神经网络的特点，使得文本深层的特征得以表示。Lee et al. (2021)的研究表明在模型构建时，同时考虑计量语言特征和深度语言特征的方法可以使模型有更好的表现。

上述可读性研究仅考虑了文本的客观特点，而阅读活动需要读者的亲身参与，文本所描写的事物、表达的感情、蕴含的哲理都会对读者产生影响。读者在阅读文本后产生的情感连接、引发的思考等也是影响阅读活动的重要因素。何应艳 (2016)提出了阅读体验的概念，认为读者在理解文字的基础上，通过阅读产生的深层次、多元的感受被称为阅读体验。阅读体验反映了作者所传达的审美意识，文本与科学美、社会美、自然美、艺术美等方面结合，体现了阅读体验的多维性，从而打破了用文本难度这一单一指标来评价文本的局限。廖圣河and 陈怡瑾 (2021)提出读者在阅读过程中的体验是沿着对文字符号、艺术形象、思想意识的深入理解逐步产生的，在调动感受力充分阅读文本得到情感、哲理、道德、人文等方面的体验后，读者的人文、科学素养和共情能力得以提升。Tracy (2017)提出阅读体验源自读者参加阅读活动的主观感受，阅读活动能够带给读者认知、情操、价值与分享四个方面的感受，因此其获得的阅读体验也分为与之对应的四种。谭继雄 (2020)则关注了读者的个体区别，认为不同的读者对阅读有着不同的需求，而需求的不同也会带来不同的阅读体验，阅读体验最终可以分为认知、审美、价值和社交四个维度。因此，我们将探究读者在阅读过程中产生的基于文本的阅读体验，把读者的阅读感受映射到文本特征上，从而使主观感受具象化，使连续属性离散化。同时，考虑阅读体验的文本在表示上也更加丰富饱满，提供了文本评价的新思路。

3 多维度阅读体验理论构建与数据建设

3.1 多维度阅读体验理论构建

由于目前对于阅读体验的研究较少，尚没有工作从多维度考虑读者在阅读过程中产生的阅读体验。因此，我们将构建一个完整的文本阅读体验的评价体系，并在此体系下完成量化阅读体验的目标。由于文本分级的重点在于篇章的分类，我们研究的对象是中文篇章，对中文篇章多维阅读体验的量化任务定义为如下形式： $G_i = \text{Classify}_i(\text{text})$ 。其中， G_i 表示某个阅读体验 i 的最终量化结果， Classify_i 指对指标 i 的分类操作。我们经过多方调查完成了阅读体验指标的选择和体系的构建。读者在阅读过程中产生的阅读体验具有多元、广泛的特点，因此我们以问卷的形式招募读者进行了广泛调查。问卷中，事先给出了可能与阅读体验相关的各种因素，读者需要阅读篇章后选择对阅读影响较大的因素。通过多方面比较和筛选，最终选定了理解性体验、文采性体验、道德性体验、思维性体验和情感性体验五个较有代表性的指标作为中文篇章多维度阅读体验指标体系的基础构成。

在阅读过程中，读者对文本的理解程度会第一时间影响阅读的体验感，过难或过简单的文本都不利于读者对文本的进一步理解，因此可理解性是必须考虑的重要指标。在我们的体系中，理解性体验按照之前可读性工作中的划分方式将小学六个年级分为三个学段，每两个年级定义为一个学段，即一二年级为学段一，三四年级与五六年级依次递增，分别标注为“0”、“1”、“2”三个类别，数字越大表示文本越不容易理解。Li et al. (2022)将文采分为“高文采”与“低文采”两个类别，我们参照该工作，以“0”表示读者对该文本的文采性体验较低，以“1”表示文采性体验较高。Wang et al. (2020)将词语的道德性分为正向道德、负向道德、中性、被动四类，在阅读文本过程中，读者会对文本中的人物形象、行为模式产生自己的道德取向，这样的道德取向性体验可以在文本中起到教育警示的作用。我们结合本次的标注语料，将文本的道德性体验分为正向、负向与中性三类，分别用“1”、“0”、“-1”表示。具有哲理性、思辨性的文本能够对读者阅读后的长期感受、价值观形成影响，我们将读者在阅读过程中感受到的关于这些方面的体验定义为思维性体验，分为两类，以“0”表示无思维相关的体验，“1”表示有思维相关的体验。读者阅读过程中感受到的不同强度的情感能够在不同程度上唤起读者的共情能力，因此情感性体验也是体系中的重要因素。我们把情感性体验分为三个类别，分别用“0”、“1”、“2”表示情感性弱、情感性适中、情感性强。最终的阅读体验类别标签情况如表1所示。

| 指标名称 | 类别数量 | 类别标签 |
|-------|------|----------|
| 理解性体验 | 3 | 0, 1, 2 |
| 文采性体验 | 2 | 0, 1 |
| 道德性体验 | 3 | -1, 0, 1 |
| 思维性体验 | 2 | 0, 1 |
| 情感性体验 | 3 | 0, 1, 2 |

表 1. 阅读体验各指标对应类别情况

3.2 多维度阅读体验数据集构建

当前没有统一的数据集用于评价读者在阅读过程中产生的阅读体验，因此我们采用人工标注的方式构建一个全新的中文篇章多维度阅读体验数据集CMRED (Chinese Multidimensional Reading Experience Dataset)。该数据集内包含篇章及其对应的多维阅读体验类别。将来，也可以通过人工标注或机器辅助的方式扩充该数据集的数量，或增加其他的指标以完善对阅读体验的探究。

我们使用骆香莹 (2022) 构建的语文教科书语料库中的小学教材作为标注语料，其中教科书语料库包含四个出版商的语文教材版本，分别为北师大版、人教版、苏教版以及义务教育课程标准实验教科书部编版，共计776篇课文。由于低年级的文本数量较少，为了保证数据基本平衡，我们又融入了绘本教材。绘本为学龄前儿童用于学习基础汉字使用的图书，文本内容较为简单，可与低年级的文本划为同一等级。我们选取了两个口碑较好、教材内容有权威性的品牌，奕阳绘本和亿童绘本，从中选择了208篇内容作为标注文本。最终总计984篇文本作为本次

标注工作的基础数据。为了确保文本长度不会显著影响阅读体验，我们以500字为准对文本进行切分。切分以句为单位，同时保证切分点位于句子末尾。最终，我们得到了1880篇由部分段落构成的不超过500字的新篇章，并随机选取其中的1200篇用于本次标注工作。

数据标注包括标注前准备和数据正式标注两个阶段。在标注前准备阶段，我们使用“锚点标注法”(唐玉玲 et al., 2022)进行本次标注，具体规则如下：对于每个指标中每一类别，提供两个篇章作为参照锚点。锚点由两名语言学专业的硕士共同商讨决定，选择最具有代表性的两个篇章作为参考标准。标注者在标注过程中与锚点进行比较，如果相似则标记相同类别，反之则考虑与锚点间比较的倾向性，是偏向更高级别还是更低级，如果介于两者之间则标记为较低等级。例如对于三分类的情感性阅读体验，判断其强度没有类别“2”那么强烈，但又比类别“1”强烈，最终仍然将该篇章归为类别“1”。这样的要求是为了统一标注上的歧义，方便后续进行结果整理。初始数据标注由五名硕士研究生完成，在正式标注开始之前，我们对标注员进行了统一的培训，明确了标注任务和标注规则，并让标注员对比锚点文本进行了试标注，为了进一步强化理解本次标注工作。在正式标注阶段，每名标注员需独立完成1200条数据的标注，即每条数据要被标注5次。标注周期为20天，每日对标注员的工作进行检查，以保证标注质量。

标注结束后，我们对五名标注员的标注结果进行了检查以及合并。对于文采性体验和思维性体验，我们按照多数投票原则选择最终类别；对于理解性体验、道德性体验和情感性体验指标，会出现选择各标签类别的标注员人数为2:2:1的情况，无法通过多数投票法选出最终类别，此时，我们让第六名标注员对该样本进行二选一的类别选择，确定标注等级。

最终我们收集到了1200条标注结果，数据集CMRED的情况如表2所示。

| 指标名称 | 理解性体验 | | | 文采性体验 | | 道德性体验 | | | 思维性体验 | | 情感性体验 | | |
|------|-------|-----|-----|-------|-----|-------|-----|-----|-------|-----|-------|-----|-----|
| 类别 | 0 | 1 | 2 | 0 | 1 | -1 | 0 | 1 | 0 | 1 | 0 | 1 | 2 |
| 文段数量 | 304 | 518 | 378 | 1045 | 155 | 39 | 859 | 302 | 1082 | 118 | 418 | 650 | 132 |

表 2. 数据集CMRED指标及其类别分布

4 方法

4.1 大规模语言模型增强的多维度阅读体验量化整体架构

当前，具备海量知识的超大规模语言模型获得了瞩目，如何利用大模型中的知识开展工作成为了一个值得研究的课题。由于我们的任务和语义知识的相关性，我们提出了基于大规模语言模型增强的多维阅读体验量化方法。多维阅读体验的某些指标之间存在相辅相成的关系，例如文采性阅读体验往往会影响读者对篇章的理解性阅读体验，因此我们使用多任务协同计算的方式来实现基础模型，我们的方法的整体架构如图2所示。

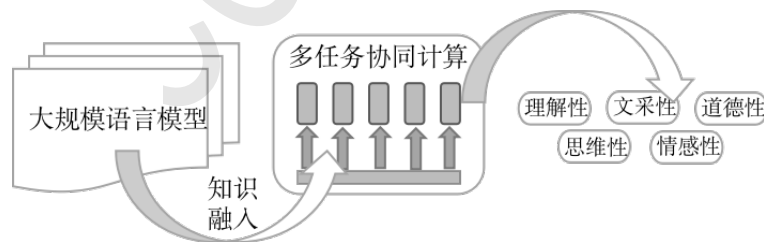


图 2. 多维度阅读体验量化整体架构

在此架构中，整个量化过程主要通过两步实现，第一步是通过合适的方式获取来自大规模语言模型的知识，第二步是将这些知识融入多任务协同计算的模型中。这两个步骤的最终输出将给读者呈现多维度的阅读体验量化结果，为读者阅读文本提供参考性指标，并能在后续的应用中实现更多有价值的工作，例如从文库中匹配与某一篇文章相似的文章或适宜某一类读者阅读的文章。

4.2 多任务协同计算方法

数据集CMRED在五个层面对篇章的阅读体验进行了指标量化，量化的过程实际上为分

类任务，最终目标是能够给出文本在各维度体验上的分类值。文本的阅读体验并非相对于文本独立存在，而是相互影响、交融，多任务学习(Crawshaw, 2020)能够合理地通过共享参数提高数据的利用效率，通过共享表示减少过拟合，因此我们使用多任务协同计算的方法来实现阅读体验量化，将5种属性的量化任务以表示共享的方式构建统一量化结构，以较小的参数量实现了模型结构。损失函数如下：

$$Loss_m = \sum_{i=1}^N \beta_i (FL_{\text{softmax}}(\text{encoder_out}))_i \quad (1)$$

其中， FL_{softmax} 表示多分类任务下的Focal Loss，Focal Loss是Lin et al. (2017)提出的用以解决样本不平衡的一种损失函数。 β 表示每项任务损失的系数，在我们的方法中，为平衡每一项分类结果，所有 β 的取值均为1。

4.3 大规模语言模型增强方法

在多维阅读体验量化体系中，一些短语能够在第一时间影响读者对某个维度的阅读体验。例如，当读者读到“伟大”、“好人好事”等表达时，会立刻触发心中对文本描述的人或事物的道德判断，随即根据文本具体内容产生更多有关道德的阅读感受。这样的表达会在一定程度上影响读者阅读下文时的关注点，也会让读者对前文内容作出短时的总结，因此我们将这种涉及到文本阅读体验的短语称为“阅读体验触发语”。阅读体验触发语是影响读者对阅读属性体验判断的关键信息，在模型中融入这些短语也会有利于模型进行量化。

ChatGPT是大型预训练语言模型InstructGPT(Ouyang et al., 2022)的后继模型，作为带有对话界面的通用语言模型，具备较强的信息抽取能力和语义理解能力。我们采用了合适的提示方式，以获取来自ChatGPT的语言知识。我们使用ChatGPT抽取指定文本的文采性、道德性、思维性、情感性相关的触发语。由于文本的理解性体验很难通过短语的方式呈现，在我们的工作中并没有标注涉及理解性体验的触发语。对于ChatGPT标注的阅读体验触发语，我们进行了预处理和人工评估，发现其标注结果和人工标注的结果基本一致。

我们共采用了三种技巧在模型中融入来自ChatGPT的知识，分别是前缀标签化提示法、标签联合监督法和标签线性化监督法。图3展示了以上三种方式在模型中融入知识的过程，整个过程分为三个阶段进行。第一阶段是利用ChatGPT进行阅读体验触发语的挖掘，第二阶段是对阅读体验触发语进行标签化，第三阶段是在基线模型中融入触发语优化，接下来我们将分别介绍以上三种基于大模型增强的阅读体验量化方法的实现过程。

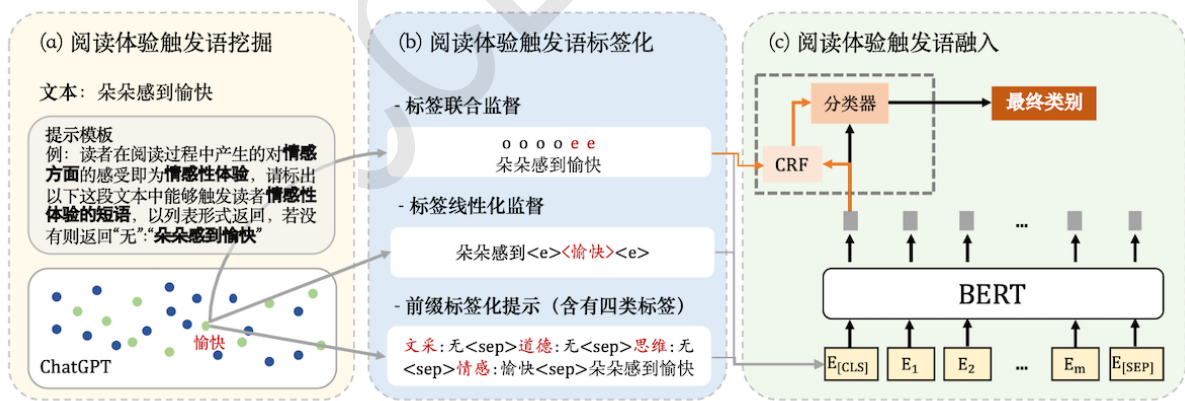


图 3. 基于大模型增强的中文文本多维阅读体验量化流程

前缀标签化提示 在提示工程的新范式(Liu et al., 2021)下，预训练语言模型的上下文学习能力可以充分体现，借鉴该思路对提示的认知，我们提出以标签化提示的方式，在模型的输入上添加提示信号。具体的形式如下：**[标签类型:标签词][原始文本输入]**。其中，标签类型包括文采、道德、思维、情感四种，输入模型时，每一种标签类型之间以分隔符号分隔。相比起提示工程从预训练语言模型中直接得到对应的输出，我们的提示方式旨在将强信号融入输入的文

本，将该信号作为模型需要学优先习的表示之一，因此需要在模型的输入中强调重点关注的标签类别和标签词语。由于该信号的强调发生在文本的初始，能起到提示信号的作用，为模型学习输入文本提供先决条件。

标签联合监督 在原始的方法中，我们用多任务学习的方法构建了量化模型的统一框架，但该框架仅在输入和指标间形成了映射关系，对文本表示的可解释性较差。对此，我们采取序列标注的方式，对文本中序列 $x_1x_2\dots x_m$ 进行基于阅读体验触发语的标注得到对应的新序列，以 g 代表文采性、 m 代表道德性、 p 代表思维性、 e 代表情感、 o 代表普通文本，即： $a_1a_2\dots a_m=\mathbf{A}(x_1x_2\dots x_m)$ 。其中 \mathbf{A} 表示标注操作。对于标签冲突的情况，我们采取随机的方式选择最终标签。在模型中，我们使用CRF(Lafferty et al., 2002)作用于基线模型的输出结果，以得到序列标注的损失。就整体模型而言，我们将此序列标注任务视作是多任务协同计算中的一环，增设一个模型通道融入多任务协同计算框架。该通道后的损失函数如下所示：

$$Loss_a = -\log\left(\sum_{t \in T} e^{score(t)}\right) - score(t_G) \quad (2)$$

$$score(t) = \sum_i^{L_t} CRF_i[t_{i-1}, t_i] \quad (3)$$

其中， $Loss_a$ 是维特比损失， t_G 表示真实的标注序列， T 表示所有可能的标注序列。 $score_t$ 定义了序列标注 t 的得分等于每个标注得分的总和，其中 L 表示序列长度。该方法增加的全序列监督信号能使阅读体验触发语以上下文位置关联的方式融入文本表示中，为后续的分类提供重点信号和位置信息。标签联合监督方法的最终损失是：

$$Loss = Loss_m + Loss_a \quad (4)$$

标签线性化监督 Ding et al. (2020)提出以线性化的方式把文本数据和序列标注变成语言模型的输入数据，从而通过语言模型的输出得到更多的数据以进行数据增强，用于解决低资源情况下的下游任务。由于序列标注的多任务学习方法以完全独立的通道实现对文本表示的监督，会造成一定的信息损失，因此我们将标签线性化的方法迁移至大规模语言模型增强中用于量化任务的表示，不同的是，我们将文采性、道德性、思维性、情感性体验的原始触发语看作是特殊标记，而其他的词看作是普通文本，将特殊标记添加在原始触发语在文本中位置的左右两端。标签线性化之后，模型的输入中既有自然语言，又有特殊标记所带来的深层属性信号，为后续的阅读体验量化任务提供完整、有效的信息。和标签联合监督法相同，我们以 g 代表文采性、 m 代表道德性、 p 代表思维性、 e 代表情感性、 o 代表普通文本，对于长度为 l 的序列 $x_1x_2\dots x_l$ ，将标签融入到该序列中，即： $m_1m_2\dots m_{l+k}=\mathbf{Linear}(x_1x_2\dots x_l)$ 。其中， \mathbf{Linear} 表示标签线性化过程， k 表示增加的阅读体验触发语标签数量，融入触发语标签后，序列长度由 l 变为 $l+k$ 。

5 实验与结果

5.1 实验设计

FewShot下对ChatGPT的能力的探究 ChatGPT是由OpenAI推出的带有图形化界面的基于大规模语料训练的自然语言处理工具(Dai et al., 2022)，我们在此大模型工具上使用FewShot方法测试了其对阅读体验指标的量化能力。

单任务独立计算 沿用传统的文本分类方法，采用预训练语言模型Bert(Devlin et al., 2019)分别对5个阅读体验指标单独建模和微调。

多任务协同计算 采用多任务学习的方法对5个阅读体验指标进行协同计算，此实验和单任务独立计算实验相同，是基于Bert模型实现的。同时，我们在此实验中使用了在第4节中提出的大模型增强方法，完成了在多任务协同计算架构下的大规模语言模型增强的多维度阅读体验量化。

5.2 实验数据与评价指标

我们将数据集CMRED中的1200条样本按照8:1:1的比例分为训练集、验证集和测试集。在实验中，epoch设置为20，学习率设置为 $1e-5$ 。为了能够准确衡量数据集类别不平衡情况下的模型性能，我们使用精确率(P)、召回率(R)和F1值(macro-F1)作为模型的评价指标。

5.3 主实验分析

ChatGPT在许多任务，例如错误信息检测、问答的FewShot结果令人惊喜(Bang et al., 2023)，但尚未有人在深层属性量化领域探索ChatGPT的能力，因此我们使用数据集CMRED探究ChatGPT在阅读体验量化上面的表现，具体的提示过程在附录A中例举。表4中是ChatGPT对CMRED测试集的FewShot结果。从表中我们可以看到，ChatGPT在各个阅读体验的量化整体上并没有获得较为准确的结果，甚至在思维性这个二分类任务上的表现较差。这可能是由于ChatGPT并不擅长以较为严苛的标准量化深层属性，而仅在上下文语义中理解文本、生成更多的文本解读，导致大规模语言模型中蕴含的知识并不能直接以量化的形式呈现。思维性作为较难定义的量化指标，更加依赖于人类的主观感受，因此ChatGPT在思维性体验指标的量化上效果更差。图4展示了针对某一个篇章，ChatGPT给出的错误思维性体验标签，可以看出，ChatGPT“有理有据”地回答了文本的思维性标签，但它在量化时受到了量化任务要求的影响，倾向于过度理解文章内容，从而导致混淆了“哲理性、思辨性”的含义。

文本X: 我国古代有一位女英雄，名叫花木兰。那时候，北方经常发生战争。一天，朝廷下达了紧急征兵的文书。木兰见到上面有父亲的名字，焦急万分。她想：父亲年老多病，难以出征；弟弟又小，还不够当兵的年龄。自己理应为国为家分忧。她说服了家人，女扮男装，替父从军。木兰告别了亲人，披战袍，跨骏马，渡黄河，过燕山，来到了前线。在多年征战中，她为国立下了赫赫战功。木兰胜利回乡后，脱下了战袍，穿上了心爱的女装。将士们前来探望她，这才惊讶地发现，昔日英勇善战的花将军，竟是位文静俊美的姑娘。

提示1: 文本中蕴含哲理内容或者思辨性内容则标注思维性体验值phi为1，若不含有则标注为0。以下是我给你的几个示例，[……]，[……]，[……]，[……]，请先学习这些示例，以完成下一步工作。

提示2: 根据学到的标注方法和规则，对以下文本标注phy值：“[文本X]”。

ChatGPT: phi为1。这篇文本表达了一个女英雄因国家需要而舍弃自己的安逸生活去从军，最终立下赫赫战功的故事。这个故事蕴含的哲理是，为国家、家庭和正义而奋斗是值得尊重和赞扬的。

图 4. ChatGPT错误标注样例

表3展现了图4中的文本的真实标签、ChatGPT预测结果和我们的标签线性化监督方法下模型的预测结果。可以看到，ChatGPT不善于分清这五种阅读体验的界限和各个量化值的含义，容易对体验值造成错误的标注，在本条标注中，对文采性、思维性和情感性的体验都产生了偏差，而我们的方法得到的预测值相对而言准确率更高。

| 模型名称 | 理解性 | 文采性 | 道德性 | 思维性 | 情感性 |
|---------|-----|-----|-----|-----|-----|
| 实际标注标签 | 1 | 0 | 1 | 0 | 0 |
| ChatGPT | 1 | 1 | 1 | 1 | 1 |
| Bert+线性 | 1 | 0 | 2 | 0 | 0 |

表 3. 模型预测样例

同时，我们采用多任务协同计算的架构实现了阅读体验量化任务，并验证了三种大模型增强的方法对于模型性能的提升效果。基于Bert的多任务协同计算方式不仅统一了整个框架从而减少了参数量，由表4我们也可以对比发现其在各维度的阅读体验量化任务上的F1值均高于单任务的结果。由此我们可以得出结论，多任务协同计算的方式比单任务独立计算的效果更好，这证实了阅读体验彼此之间存在的正向影响。

| 模型名称 | 理解性 | 文采性 | 道德性 | 思维性 | 情感性 |
|---------|----------------|----------------|----------------|----------------|----------------|
| | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 |
| ChatGPT | 0.56/0.45/0.40 | 0.68/0.76/0.69 | 0.45/0.50/0.42 | 0.58/0.70/0.43 | 0.48/0.48/0.45 |
| Bert单任务 | 0.70/0.69/0.68 | 0.74/0.74/0.72 | 0.42/0.47/0.44 | 0.84/0.62/0.67 | 0.49/0.52/0.50 |
| Bert多任务 | 0.73/0.73/0.72 | 0.84/0.74/0.77 | 0.48/0.52/0.49 | 0.77/0.73/0.75 | 0.55/0.52/0.53 |
| Bert+提示 | 0.77/0.75/0.75 | 0.78/0.74/0.76 | 0.49/0.54/0.51 | 0.84/0.62/0.66 | 0.66/0.59/0.60 |
| Bert+联合 | 0.71/0.71/0.71 | 0.87/0.80/0.83 | 0.89/0.62/0.67 | 0.45/0.50/0.47 | 0.63/0.56/0.68 |
| Bert+线性 | 0.79/0.78/0.78 | 0.86/0.86/0.86 | 0.81/0.60/0.60 | 0.78/0.69/0.73 | 0.75/0.60/0.64 |

表 4. 模型在各维度阅读体验上的分类表现

从表中不难发现，不论采取哪种策略，融入ChatGPT的知识的多任务协同计算方法的效果都在一定程度上比多任务协同计算的基础模型的效果更好。其中，标签联合监督法和标签线性化监督法分别在不同类别的量化任务中表现优于前缀标签化提示法，我们认为这可能是由于前两者以不同方式实现了整个文本在对应位置上的标记融合，这能够更好地帮助模型学习上下文中的深层属性，而前缀标签化提示法仅给予了输入信号，却缺乏标签的位置信息，使得模型无法学习到标签和自然语言之间的对应位置关系。另外，实验结果表明，在采用大模型增强方法的情况下，尽管多数体验的量化效果有所提升，但思维性体验的量化效果却比只采用多任务协同计算架构的效果要差。图5展示了基础的多任务架构和标签线性化监督方法在各维度指标上的F1值，可以看到除了思维性体验指标，其他指标均在线性方法上得到了F1值的提升，我们将在下一小节中讨论这一点。

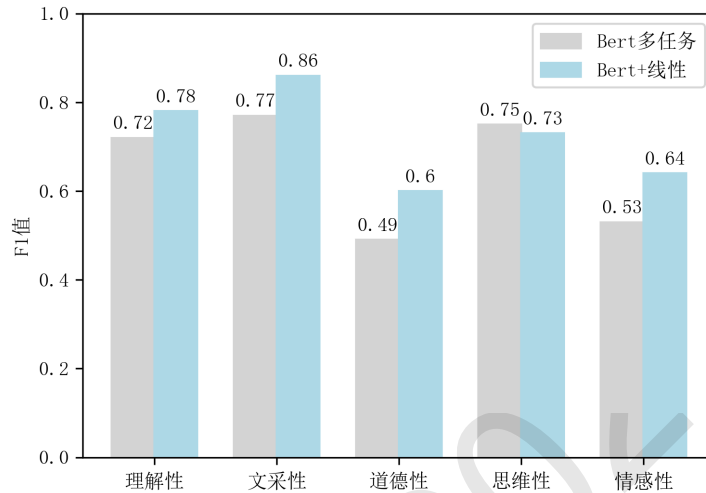


图 5. 多任务基础架构和标签线性化方法下各指标结果的F1值

5.4 探究实验分析

(1) 大模型增强对思维性体验量化的影响

基于ChatGPT增强的方法降低了多任务协同计算架构下对于思维性阅读体验的量化效果。对此，我们推测可能是思维性阅读体验的抽象性导致其在量化上更难实现，且由于该指标具备相对的独立性，对思维性的特征学习需要借助文本整体的上下文含义。因此，加入的标签越多、对其他体验的学习越深入，对思维性体验的学习越容易造成混淆。

我们在多任务协同计算的架构下对思维性阅读体验量化展开了单独的研究，在标签线性化监督法上分别探究了仅添加文采性触发语、道德性触发语、情感性触发语或思维性触发语本身的情况下思维性阅读体验的表现，表5展示了该实验下思维性阅读体验的计算结果。从表中我们可以发现，文采性阅读体验、道德性阅读体验标签对思维性阅读体验计算的结果产生了负面影响，这可能是由于文采性和道德性阅读体验标签同思维性阅读体验的联系较少，添加这两类标签导致了多余噪声的产生。再者，思维性体验作为一种具备开放性的属性，是一种无限集合，然而文采性和道德性相关的触发语能够在一定程度上被划分为有限的几种类型，有限集合的知识融入对无限集合相关的指标量化有损。

| 模型名称 | P | R | F1 |
|--------|------|------|------|
| 全部标记 | 0.78 | 0.69 | 0.73 |
| 仅文采性标记 | 0.76 | 0.62 | 0.65 |
| 仅道德性标记 | 0.71 | 0.61 | 0.64 |
| 仅情感性标记 | 0.77 | 0.73 | 0.75 |
| 仅思维性标记 | 0.71 | 0.61 | 0.64 |

表 5. 其他触发语在标签线性化监督法中对思维性阅读体验的影响

此外，我们发现思维性标签本身也易对结果产生负面的影响，这可能是由于思维性阅读体验非常依赖于长距离的上下文判断，导致标注的阅读体验触发语具有多样性、特殊性的特点，因此此类标签的添加不利于模型的泛化性。图6是ChatGPT标注的思维性体验触发语的示例，可以看到，思维性体验触发语和文本本身的相关性极强且跨度较大，ChatGPT对思维性的理解较为灵活，在两个不同的文本中标注的表达完全不同。

标注1: 无论如何，我不能使家乡的孩子失望，我终于拿起了笔。请原谅，我今年不能回家乡，并不是不愿意看望你们，正相反，我多么想看见你们天真的笑脸，多么想听见你们歌唱般的话语，但是我没有体力和精力支持这样一次长途的旅行。那么，就让这封信代替我同你们见面吧。不要把我当作什么杰出人物，我只是一个普通人。我写作，不是我有才华，而是我有感情，对我的祖国和同胞有无限的爱，我用作品表达我的这种感情。我今年八十七岁，我**思索**，我**追求**，我终于明白**生命的意义**在于奉献而不在于享受。我在回答和平街小学同学们的信中：“我愿意再活一次，重新学习，重新工作，让我的生命开花结果。”

标注2: 认识大自然文字的人，立即会说，它不是人搬来的，而是冰搬来的。那些冰块从寒冷的北方“爬”过来，沿路把大大小小的石块带着一起走。这是好久好久以前的事了，当时这儿根本就没有森林。周围的森林是后来才长起来的。要学会**认识大自然的**文字****，从小就应当到树林里或者田野上走走，注意观察。假如有什么不明白的地方，应再到书里去寻找，看那里有没有解释。你还应该去请教有学问的人：这是什么石头？这是什么树？总是**坐在家里的人，永远不会懂得大自然的**文字****。”

图 6. ChatGPT对思维性体验的标注样例

(2) 思维性体验量化任务和其他任务之间的关系

我们在探索中发现了思维性体验的特殊性，其特殊性不仅在于知识表示形式上的特殊，更在于体验形成过程的特殊。我们认为思维性体验同其他任务不同，需要进行跨度较大的上下文学习与总结，因此我们探索了思维性体验量化任务和其他任务之间的关系。我们分别用多任务协同计算基础架构和标签线性化监督方法实现了去除思维性量化任务的实验，表6是实验结果的展示。

从表中我们可以看出，去除思维性任务后，其他任务的F1值有所提升。但在多任务协同计算的基础架构下，思维性任务的效果要好于单独建模思维性的量化效果。实验结果表明，与其他任务的联合学习对思维性任务的量化有一定的帮助，但思维性任务会造成其他任务量化性能的下降。这说明思维性体验除了在知识上与其他任务不同，在表示上也和其他任务有异，该任务需要在长上下文中学习对应的总结性特征，从而对其他体验的量化产生干扰。而其他任务的知识虽然与思维性无关，但量化过程中能够产生的多维度总结性表示，对思维性体验的量化有益。

| 模型名称 | 理解性 | 文采性 | 道德性 | 思维性 | 情感性 |
|---------|----------------|----------------|----------------|----------------|----------------|
| | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 |
| Bert单任务 | - | - | - | 0.84/0.62/0.67 | - |
| Bert多任务 | 0.73/0.73/0.72 | 0.84/0.74/0.77 | 0.48/0.52/0.49 | 0.77/0.73/0.75 | 0.55/0.52/0.53 |
| 无思维性 | 0.79/0.78/0.78 | 0.86/0.86/0.86 | 0.52/0.52/0.52 | - | 0.64/0.58/0.60 |
| 线性+无思维性 | 0.78/0.79/0.76 | 0.88/0.87/0.87 | 0.67/0.67/0.67 | - | 0.76/0.60/0.62 |

表 6. 去除思维性体验任务后各维度体验上的分类表现

6 结论

针对中文文本分级和读物推荐中存在的标准单一、体系缺失的现状，我们提出了中文篇章多维度阅读体验量化体系，以完善中文文本分级及推荐的依据和标准。基于此，我们构建了中文篇章多维度阅读体验数据集CMRED。我们又提出了三种融入ChatGPT知识的方式来实现基于大模型增强的阅读体验量化方法，分别是前缀标签化提示法、标签联合监督法和标签线性化监督法。这三种增强方式在后续实验中的表现均超出了基线模型，证明了增强方式有效性的同时，也说明了大模型知识对于阅读体验量化任务的正向作用。同时，我们使用构建的体系及数据集探究了ChatGPT在阅读体验量化上的表现，发现ChatGPT更擅长自然语言的理解和生成，但不善于对深层的特征进行量化。未来，我们将会继续探索更多阅读体验的指标，同时在基于大模型增强的方法上作出其他尝试，而对于探究实验中发现的思维性指标特殊性的持续研究，也是我们未来工作的重点。

参考文献

- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.
- Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey.
- Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers.
- Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online, November. Association for Computational Linguistics.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
- J. Lafferty, A. McCallum, and Fcn Pereira. 2002. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *proceedings of icml*.
- Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.
- Yi Li, Dong Yu, and Pengyuan Liu. 2022. CLGC: A corpus for Chinese literary grace evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5548–5556, Marseille, France, June. European Language Resources Association.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.
- Danielle S McNamara, Max M Louwerse, Philip M McCarthy, and Arthur C Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.
- Simin Rao, Hua Zheng, and Sujian Li. 2021. 阅读分级相关研究综述(a survey of leveled reading). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 689–702, Huhhot, China, August. Chinese Information Processing Society of China.
- Karen A Schriver. 2000. Readability formulas in the new millennium: what’s the use? *ACM Journal of Computer Documentation (JCD)*, 24(3):138–140.
- Dean R Smith et al. 1989. The lexile scale in theory and practice. final report. *analysis of variance*.
- Daniel G Tracy. 2017. Libraries as content producers: How library publishing services address the reading experience. *College & Research Libraries*, 78(2):219.

- Hongrui Wang, Chang Liu, and Dong Yu. 2020. 面向人工智能伦理计算的中文道德词典构建方法研究(construction of a Chinese moral dictionary for artificial intelligence ethical computing). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 539–549, Haikou, China, October. Chinese Information Processing Society of China.
- Siyuan Wu, Jianyong Cai, Dong Yu, and Xin Jiang. 2018. A Survey on the Automatic Text Readability Measures(文本可读性的自动分析研究综述). *中文信息学报*, 32:1–10, December.
- 何应艳. 2016. 小学语文阅读教学中美育的实施. 硕士, 云南师范大学.
- 刘苗苗, 李燕, 王欣萌, 甘琳琳, and 李虹. 2021. 分级阅读初探:基于小学教材的汉语可读性公式研究. *语言文字应用*, (2):116–126.
- 唐玉玲, 张宇飞, and 于东. 2022. 结合深度学习和语言难度特征的句子可读性计算方法. *中文信息学报*, 36(2):29–39.
- 廖圣河and 陈怡瑾. 2021. 小学生语文阅读感受力的构成和培养. *语文建设*, (20):19–23.
- 杨春and 张秋月. 2022. 文本阅读体验应是语文课堂教学的发端与根本. *中华活页文选:教师*, (7):3.
- 池春红. 2022. 激发阅读兴趣,增强阅读体验,促进深度学习——初中语文名著阅读教学的有效性探究. *亚太教育*, (21):91–94.
- 程勇, 徐德宽, and 董军. 2020. 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究. *语言文字应用*. *语言文字应用*, (1):132–143.
- 谭继雄. 2020. 公共图书馆阅读体验的影响因素研究——基于青少年群体阅读风格视角. *图书馆学刊*, 42(8):68–72.
- 骆香莹. 2022. 基于多元语言特征与眼动特征的中文可读性自动评估. 硕士学位论文, 北京语言大学.

A 思维性体验的提示模板和ChatGPT响应

提示1: 文本中蕴含哲理内容或者思辨性内容，能引发读者哲思的，则标注思维性体验值 ϕ 为1，若不含有则标注为0。以下是我给你的几个示例，示例1: [文本: 这时候，我清清楚楚听见老师说，按照进化论的观点，母腹中的胎儿再现了从简单生命进化成人的过程。当时教室里安静得出奇，大家都默不作声。我忽然想起了自己的发现，情不自禁地笑出了声音。老师狠狠地瞪了我一眼。“费奥多罗夫！……你笑什么？再笑就从教室里出去！”“奥尔加·伊万诺夫娜，我……我想起了自己的发现……”教室里一阵笑声。奥尔加伊万诺夫娜气得脸色苍白，大步朝我走来。“费奥多罗夫！……你立刻从教室里出去！……”我的脸由于困窘和羞愧一下子涨得通红。这时候我意识到，老师误解了我的笑声，以为我的笑不怀好意。幸亏她没有容我解释，不然的话，同学们听见我说自己三年前就发现了进化论，还不笑塌房顶！不过，被轰出教室，站在外面，我倒想出了一条和自我安慰的理由，我明白了——世界上重大的发明与发现，有时还面临着受到驱逐和迫害的风险。此文本的 ϕ 值为1。]，示例2: [文本: 春雨，像春姑娘纺出的线，轻轻地落到地上，沙沙沙，沙沙沙……田野里，一群小鸟正在争论一个有趣的问题：春雨到底是什么颜色的？小燕子说：“春雨是绿色的。你们瞧，春雨落到草地上，草就绿了。春雨淋在柳树上，柳枝也绿了。”麻雀说：“不对，春雨是红色的。你们瞧，春雨洒在桃树上，桃花红了。春雨滴在杜鹃丛中，杜鹃花也红了。”小黄莺说：“不对，不对，春雨是黄色的。你们看，春雨落在油菜地里，油菜花黄了。春雨落在蒲公英上，蒲公英花也黄了。”春雨听了大家的争论，下得更欢了，沙沙沙，沙沙沙……此文本的 ϕ 值为0。]，请先学习这些示例，以完成下一步工作。

ChatGPT回答: 根据您提供的示例，示例1中的文本涉及到思辨性内容和哲理思考，因此可以标注为思维性体验值(ϕ)为1。这是因为文本描述了主人公对自己的发现以及对被误解的感受的思考，涉及到了对人类发明与发现所面临的风险和迫害的理解。示例2中的文本主要描述了小鸟们对春雨颜色的争论，没有明显的哲理或思辨性内容，因此可以将其标注为思维性体验值(ϕ)为0。这段文本更多地展现了小鸟们之间的对话和辩论，而不是探讨深层次的思考或哲学观点。根据上述解释，示例1的 ϕ 值为1，示例2的 ϕ 值为0。

提示2: 根据学到的标注方法和规则，对以下文本标注 ϕ 值：东汉的时候，有个青年叫陈蕃，他虽志存高远，喜好读书，但生活上却很懒散，连自己的书房也很少收拾打扫。有一天，一位叫薛勤的人来拜访陈蕃的父亲，正巧他父亲外出不在，只有陈蕃在家。陈蕃见来了客人，就从自己的书房里出来接待，告诉客人父亲现在不在家。没料到薛勤一边与他寒暄，一边随意地走进了陈蕃的书房。薛勤一进书房，着实吓了一跳。屋子里乱七八糟，桌子上书呀，杂物呀，什么都有，上面满是灰尘；地上到处是垃圾；墙壁止有不少蜘蛛网。屋里不仅肮脏，空气也很污浊。薛勤看到后，不禁皱起了眉头。他对陈蕃说：“年轻人，你的屋子怎么弄得这么乱？为什么不把屋子打扫干净呢？”没想到陈蕃却满不在乎地说：“大丈夫活在世上，要干的是轰轰烈烈的大事业，要扫除的是天下一切不平之事，哪里会花心思去清扫小小的一间屋子呢？”他说得理直气壮，很有些得意的神气。听着陈蕃这样的回答，薛勤心里暗想：此人虽年少而有壮志，但连小事都不愿意做，又怎么能做成大事呢？于是他盯着陈蕃，反问道：“年轻人，你连一间小小的屋子都不扫，又怎么去扫天下呢？”薛勤走后，陈蕃沉思起来，觉得他的话很有道理。

ChatGPT回答: 根据您的示例，我将对文本进行标注。这篇文本的 ϕ 值为1。这篇文本含有哲理，通过薛勤的质问，告诫读者做事要有条理，不要忽略小事。