提示词语气对国产大语言模型输出结果的影响分析*

王兵书[1] 郑又祥[1] 李佳乐[1] 郑江滨[1] 王明明[2] 张晗[3**]

（1.西北工业大学 软件学院，陕西 西安 710129；
2.海南师范大学 教师教育学院，海南 海口 571158；
3.西北工业大学 光电与智能研究院，陕西 西安 710072）

摘 要 提示词语气是大语言模型使用中常见的输入差异，但其影响并不只体现在最终答案上。本文以 2023—2025 年浙江卷纯文本选择题作为标准化教育测评任务，面向 DeepSeek、Doubao 和 Qwen 三种国产大语言模型，构建统一的提示词语气测试框架。在保持题目内容和输出格式一致的条件下，仅调整提示词的礼貌程度，并从准确率、词元成本和显性作答结构三个方面比较模型响应。实验结果表明，不同语气未形成简单的线性准确率差异；在本文样本中，礼貌语气是唯一在三种模型上均未低于中性语气的语气。与正确率相比，词元成本和回答结构对语气变化的响应更为明显，且变化方向依赖具体模型。本文结果可为提示工程课程、大模型通识课和学生 AI 工具使用规范提供实验依据。

关键词 提示词语气；大语言模型；计算机教育；提示工程；计算成本；回答结构

Analysis of the Impact of Prompt Tone on the Outputs of Chinese Large Language Models*

Bingshu Wang[1], Youxiang Zheng[1], Jiale Li[1], Jiangbin Zheng[1], Mingming Wang[2], Han Zhang[3**]

1. School of Software, Northwestern Polytechnical University, Xi'an 710129, China;
2. School of Teacher Education, Hainan Normal University, Haikou 571158, China;
3. Institute of Optics and Intelligence, Northwestern Polytechnical University, Xi'an 710072, China

Abstract: Prompt tone is a common input variation in the use of large language models, but its influence is not limited to final answer accuracy. Using text-only multiple-choice questions from the 2023–2025 Zhejiang examinations as standardized educational assessment tasks, this study constructs a unified prompt-tone testing framework for three Chinese large language models: DeepSeek, Doubao and Qwen. With question content and output format held constant, only the politeness level of the prompt is varied. Model responses are compared in terms of accuracy, token cost and visible answer structure. The results show no simple linear relationship between tone and accuracy. Within the tasks examined, `Polite` is the only condition that does not underperform `Neutral` on any of the three models. Compared with accuracy, token use and visible answer structure respond more clearly to tone, although the direction of change remains model-dependent. The results provide experimental evidence for prompt-engineering courses, general education on large language models and student guidelines for AI-tool use.

Keywords: prompt tone; large language model; computer education; prompt engineering; token cost; answer structure

***基金资助：**本文得到西北工业大学教育教学改革研究项目(重点)(项目编号:2026JGZ057)、西北工业大学教育教学改革研究(项目编号:2025JGZY55)基金资助、陕西省学位与研究生教育研究面上项目(项目编号:SXGERC2026031)和海南省哲学社会科学规划课题(项目编号:HNSK(ZC)25-291)基金资助。

1 引 言

近年来，大语言模型在中国的应用规模持续扩大。国家数据局2026年5月发布的数据显示，2025年全国词元调用量约21100万亿，日均调用量由年初的超万亿增长至年末的100万亿[1]。在这种背景下，用户在解决信息检索、内容生成、教育学习和辅助决策等问题时，越来越多地会借助提示词与大语言模型进行交互。提示词通常是指用户输入给模型的自然语言指令，提示工程则是在此基础上围绕提示词进行设计、优化和迭代的过程。已有研究表明，提示词的组织方式会改变模型的推理路径和输出表现，规范的提示设计可以提高结果质量和稳定性[2][3]。

在提示词设计因素中，语气和礼貌程度是一类比较容易直接调整的表述变量。已有研究表明，提示词礼貌程度会影响模型表现，但这种影响并不呈现稳定的线性规律，还会随着模型种类、任务类型和语言场景而变化[4-6]。同时，提示词表述方式和语言风格的细微变化，还可能改变模型的推理过程与输出成本[3][7][8]。这说明，判断提示词语气的作用，不能只看最终答案的高低，还要把模型生成过程中的资源投入和回答组织方式一并看进去。

这一问题在计算机教育场景中也有直接含义。学生已经把大语言模型用于资料检索、代码理解、程序调试、课程作业辅助和学习反馈，教师也开始在提示工程课程、大模型通识课和 AI 辅助编程教学中讲授提示词设计。若提示词语气主要改变词元成本和作答结构，而不是稳定提高准确率，那么课堂中的提示词教学就不能只强调“更礼貌”或“更强硬”，还需要把成本意识、解释结构和适用边界纳入学生 AI 工具使用规范。

基于这些考虑，本文面向DeepSeek、Doubao和Qwen三种国产大语言模型，提出了一套统一的提示词语气测试框架。在保持题目内容和输出格式一致的条件下，本文只改变提示词的礼貌程度，并同步测量准确率、词元成本和显性作答结构，从而判断语气效应是先出现在结果上，还是先出现在过程上。和已有主要围绕准确率展开的礼貌语气研究相比[4-6]，本文把成本和回答结构也放进了同一分析框架。该研究可以为中文提示工程课程、大模型通识课和学生AI工具使用规范提供经验依据，也可以为标准化客观题中的提示词对照实验提供任务素材。

2 提示词影响大模型输出结果的相关研究

提示词与模型表现之间的关系，已经有大量研究从不同角度展开讨论，但把语气、成本和结构放在同一框架里一起考察的工作仍然不多。本节从礼貌与语气、表述方式与推理、过程性推理以及评测可靠性四个方面梳理既有工作，并交代本文与这些研究之间的关系。

2．1 提示词礼貌与语气研究

关于提示词礼貌与模型表现之间关系的研究已经初步表明，粗鲁提示往往会降低模型表现，但过度礼貌也不一定最优，礼貌效应并不呈简单的线性梯度[4]。围绕提示词礼貌度的近年复现实验以及跨模型评估也显示，礼貌效应会随着模型代际、任务类型和模型家族而变化，未必都会朝同一个方向提升或下降[5][6]。Liu在逻辑任务上的比较结果进一步表明，提示词的具体性和表述方式会系统性地改变模型表现，而且这一效应在通用模型与推理强化模型之间存在差异[7]。这些工作说明，语气并不只是表面的修辞外壳，它还可能影响模型如何理解当前任务情境。本文在此基础上进一步聚焦中文场景，并把分析对象扩展到准确率、计算成本和回答结构三个方面。

2．2 提示词表述方式与推理表现

除了礼貌度之外，提示词的表述方式本身也会改变模型的推理表现。已有研究表明，即使任务语义保持不变，提示词表述方式的细微差异仍可能稳定改变模型输出[3]。从实际部署成本来看，有研究指出提示词中的语言风格变化本身就可能改变输出词元数量；即使回答质量没有同步变化，也会放大实际应用中的成本不确定性[8]。此外，也有研究从注意力分散和响应失稳的角度，尝试解释提示词变化为什么会影响模型行为[9]。这说明，提示词研究不应只停留在“说了什么”，还要关注“以什么方式说”；语气研究正是在这一更广义的提示框架框架中展开的。

2．3 过程性推理与回答结构

多步推理研究已经反复表明，最终答案之外的过程信息同样有分析价值。显式推理链能够改变模型的任务表现[2]，自校验与逐步验证则提示复核过程本身可能构成独立的能力维度[10-11]。进一步看，已有工作还尝试把多步推理拆分为路径生成和答案校准两个阶段[12]，并区分过程反馈与结果反馈的不同作用[13]。这些研究为本文分析任务复述、正式推理、选项分析、检查总结和最终答案之间的资源分配提供了理论背景。

2．4　大模型评测与Benchmark可靠性

在大模型评测研究中，数据真实性、题目来源可靠性和Benchmark污染问题，会直接影响实验结论的解释力[14]。如果评测数据本身存在来源不清、答案不稳或样本污染问题，那么模型差异和语气差异都可能被高估或误读。因此，本文在实验之前对题目来源、答案一致性和样本可用性做了较细致的人工核验与清洗。

综合以上四个方面可以看出，目前的研究在礼貌语气、推理过程和评测规范上各自都有一定积累，但把这三者整合起来，并在统一的中文客观题场景中进行联合分析的工作仍然较少。本文正是在这一交叉点上展开：既控制评测数据的可靠性，也同时观察最终答案、词元成本和回答结构三个方面的语气效应。

3　提示词语气对国产大模型输出结果影响的设计方法

本节说明本文的研究框架、变量设置、回答结构编码方法以及统计比较口径。围绕提示词语气这个核心自变量，本文同时考察准确率、计算成本和回答结构三个方面，用来识别语气变化在模型输出中会表现在哪里。

3．1　整体研究思路

图1展示了本文提示词语气实验的整体方案和分析路径。整个研究自上而下分为输入层、处理层和分析层。输入层负责固定题目内容、构造五档语气并统一模型调用条件；处理层负责批量运行、结果提取和数据清洗；分析层则分别开展准确率分析、计算成本分析、回答结构分析以及错误题目重分析。

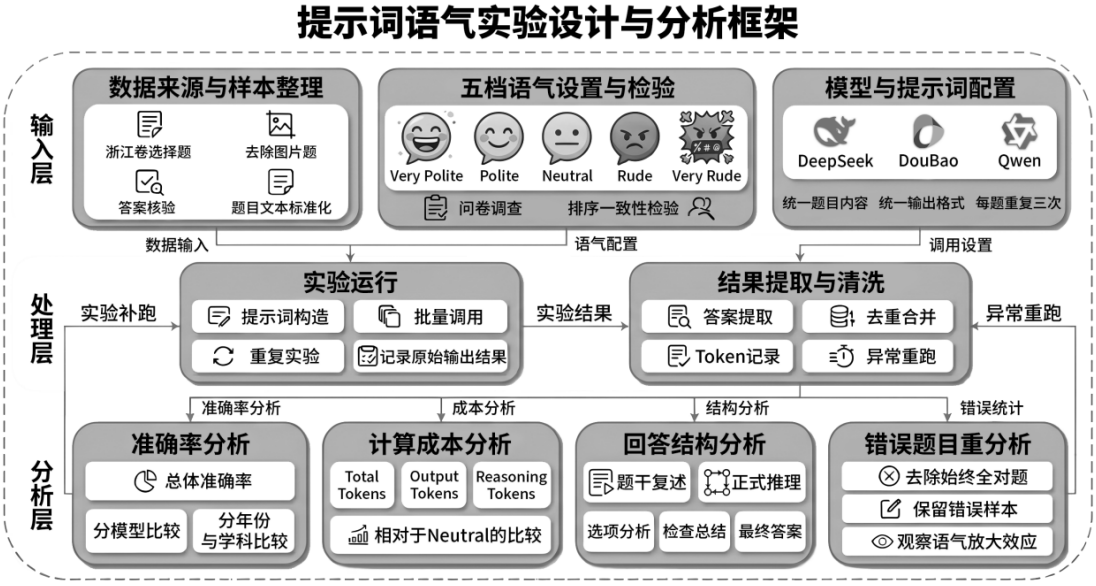

图 1　提示词语气实验的研究设计与数据分析流程

在这一框架下，提示词语气作为输入变量进入统一实验流程，随后再从最终作答结果、词元资源投入和显性回答组织方式三个方面观察模型响应。这样一来，语气效应既可以在最终答案上比较，也可以在作答过程上加以识别。

3．2　变量设计与核心指标

本文的核心自变量是五档提示词语气：非常礼貌、礼貌、中性、粗鲁和非常粗鲁。五档提示词共用同一题目内容、同一回答格式和同一输出要求，变化仅体现在礼貌措辞、态度强度以及语气强弱上。

围绕这一自变量，本文设置了三类核心指标。第一类是准确率，用来衡量模型最终作答与标准答案的一致程度。第二类是计算成本，主要考察总词元数、输出词元数、推理词元数以及相对中性语气的词元变化。第三类是回答结构，用来观察模型可见输出中的功能片段构成，以及不同语气下各类片段比重的变化。

三类指标分别对应结果质量、资源消耗和回答组织方式三个观察方面。把它们放在同一分析框架里，可以更清楚地区分语气差异是先出现在结果上，还是先出现在过程上。

3．3　回答结构分类方法

回答结构分析是本文的方法重点之一。对模型已经显性输出的文本，本文先按换行和句末标点进行切分，再依据显性词语、片段位置以及答案格式进行功能编码。位于回答前部且包含“题干”“题目”“材料”“已知”“根据题意”等提示词的片段，记为题干复述类；显式讨论A/B/C/D等选项内容，或包含“排除”“选项A”“选项B”等标记的片段，记为选项分析类；包含“综上”“因此”“故”“再检查”“唯一”等收束性标记，并且通常位于回答后部的片段，记为检查总结类；显式给出答案的片段，如“答案：A”，记为最终答案类；未命中前述规则但仍承担解释功能的片段，则归入正式推理类。

此外，本文还保留显性社交表达和拒答/安全声明两个补充类别，用来处理少量特殊输出。这样分类，是为了把回答中的不同功能片段分开，再判断语气变化主要把新增内容推向复述、分析、检查还是答案收束。

当同一片段同时命中多个规则时，本文按“最终答案、拒答/安全声明、选项分析、检查总结、显性社交表达、题干复述、正式推理”的顺序判定优先级。完成分类后，再按片段长度比例把这一通道中的词元分配到各功能类别上，据此统计不同语气条件下的显性作答片段分布。

为降低回答结构编码的主观性，本文在规则脚本之外增加了人工抽样复核。具体做法是：按模型、语气、通道和自动标签分层抽取 240 条片段，由两名标注者先依据统一准则独立标注，初始一致率为 67.08%，Cohen’s κ=0.5820；随后对分歧样本逐项复核并形成最终共识标签。以该共识标签检查并校正规则脚本后，当前脚本与共识标签的一致率为 77.92%（187/240）。因此，本文把回答结构分类作为规则辅助的功能片段指标，用于比较不同语气下的分布方向和相对变化；它不被解释为对每句话深层语义或隐式推理质量的人工金标准。由于词元分配依据片段长度比例估计，相关结果主要服务于跨条件比较，而不是单条回答的精确词元归因。

3．4　统计比较与证据判定

本文从整体、模型、年份场次以及学科等多个口径展开比较。准确率差异主要借助分类统计检验来衡量，计算成本差异主要借助均值比较和方差分析来衡量，回答结构差异则借助各类显性作答片段占比的变化来呈现。其中，相对中性语气的词元指标用来控制题目和模型的基线差异，以比较同一道题在不同语气下的资源投入变化。显著性判断以$p<0.05$为参考阈值，但本文不把单一p值作为充分证据，而是同时结合效应方向、模型一致性和困难样本重分析进行解释。

除直接统计量外，本文还使用了两个描述性指标。一是平均结构分布偏移，指五个核心片段占比相对中性语气的平均绝对偏离；二是分析—检查差值，指检查总结类片段占比变化减去选项分析类片段占比变化，用来概括新增内容更偏向逐项分析，还是更偏向后段检查。

本文还引入“错误题目重分析”作为辅助比较。具体做法是：以“模型+ 基础题目”为单位，剔除在全部语气与重复条件下始终全对的题目，只保留至少出现过一次错误的题目，并在这些困难样本上重新统计各语气的准确率。这里的重复轮次固定为3次。这样做，是为了观察语气差异在更有区分度的困难样本中是否更明显。

在结果表述上，本文先报告可以直接核验的统计结果，如整体卡方检验、按模型配对的McNemar检验、Cochran’s Q检验、单因素ANOVA、Kruskal–Wallis检验、平均词元差值以及各类片段占比变化；随后再使用描述性指标说明不同结果之间的关系。其中，整体卡方检验的统计量写为：

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \tag{1}$$

其中O_{ij}与E_{ij}分别表示第i行第j列单元格的观测频数和期望频数。单因素ANOVA的统计量为：

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/(g-1)}{SS_{within}/(N-g)} \quad (2)$$

其中g表示组数，N表示样本量。按这一口径处理后，方法说明、结果报告和讨论解释可以分开表述。

4 实验数据采集和参数设置

在前一节的方法设计基础上，本文把研究放到统一的中文客观题场景中来做。下文依次交代样本来源、五档语气文本、模型调用条件以及正式运行时的数据整理方式。

4．1 数据来源与样本整理

本文数据来自2023—2025年浙江卷，属于标准化教育测评材料。试题来源明确、答案唯一且便于稳定核验，适合在固定题目内容和输出格式的条件下开展提示词语气对照实验。数据覆盖1月与6月试卷，涉及语文、数学、英语、物理、化学、生物、政治、历史和地理九门学科。研究对象限定为适合纯文本输入、标准答案唯一且能够稳定核验的选择题，含图题、题干不完整题以及答案无法稳定核验的题目都被排除。经过筛选和人工核对后，最终保留635道能够在统一提示条件下进行批量实验的标准化题目样本，其中2023年210道、2024年220道、2025年205道。按照五档语气展开后，共形成3175组“题目—语气”实验输入；再考虑三次重复和三种模型，最终纳入正式统计的记录数为28575。

为提高实验可信度，本文对题目文本做了进一步清洗，包括核验标准答案、补全材料题题干、统一材料题拆分规则、清除图片依赖内容，以及把部分公式和特殊符号转换为稳定文本格式。完成这一步之后，原始真题才被整理成适合大语言模型批量调用的标准化样本。

4．2 五档语气文本与感知检验

本文设置五档语气：非常礼貌、礼貌、中性、粗鲁和非常粗鲁。五档提示词在任务信息、回答格式和输出要求上保持一致，只调整礼貌措辞、态度强度和语气强弱。这样设置以后，研究方法中定义的语气变量就变成了可直接运行的正式提示词。

在进入正式实验前，这五条中文表达先做了一次受试者感知检验。问卷要求受试者按礼貌程度对五条候选表达进行排序，再换算为礼貌分值。本文采用的换算方式为

$$Score = 6 - Rank \quad (3)$$

其中$Rank = 1$对应最礼貌，$Rank = 5$对应最不礼貌。这样处理后，分值方向与“*非常礼貌语气*”到“*非常粗鲁语气*”的语气顺序保持一致。问卷共获得59份有效答卷。结果显示，五条表达的平均礼貌分值分别为4.92、4.08、2.86、2.14和1.00，呈严格递减趋势，与本文预设的五档语气顺序一致；相邻档位的95%置信区间没有重叠，说明五类表达之间有清楚的层级区分。按完整排序结果计算得到的Kendall’s W= 0.961，表明受试者之间具有很高的一致性，其中47/59份答卷给出了与预设顺序完全一致的五档排序。

4．3 模型版本、重复设置与清洗规则

实验对象为DeepSeek、Doubao和Qwen。本文调用的模型版本分别为DeepSeek-V3.2-Reasoner（deepseek-reasoner）、Doubao-Seed-2.0-Pro（doubao-seed-2-0-pro-260215）和Qwen3.6-Plus（qwen3.6-plus）。每道基础题在五档语气下都运行3次，因此每个模型形成9525条正式记录。实验过程中统一了题目信息和输出格式，并尽量把模型外部调用条件保持一致，以减弱非语气因素带来的额外波动。

模型调用参数也保持统一。正式实验中，temperature 设为 0.0，max_tokens 设为 32768，单次请求超时时间为 900 秒，并发数为 25，请求抖动为 1.0 秒；对支持推理强度参数的接口，reasoning_effort 设为 medium。上述设置使重复运行更接近确定性条件，同时也为推理词元和总词元的比较保留一致的上限口径。

运行日志同时保留模型回答、答案提取结果、接口返回的词元用量以及异常重跑记录。正式统计前，原始记录按“去重—补跑—剔除”的顺序整理：同一题目在同一语气和同一轮次下如果出现重复请求，仅保留与实验设计一致的正式记录；如果记录存在返回不完整、答案提取失败或词元用量缺失等异常，则按既定规则补跑或剔除。整理完成后，三模型均保留9525条正式记录，且词元用量信息覆盖完整，因此主分析直接使用接口返回的词元统计结果。

5 提示词对大模型输出结果的分析

5．1　语气与正确率

图2给出了三种模型在整合口径下相对中性语气的总体变化。礼貌语气是唯一在三种模型上均未低于中性语气的语气：DeepSeek为+0.157个百分点，Doubao为+0.000个百分点，Qwen为+0.420个百分点。在本文模型和中文客观题样本中，这一结果表现出较一致的描述性稳定性，但尚不能认为礼貌语气相对中性语气具有统计显著优势，更不能据此推出“语气越礼貌，回答越准确”的结论。整体上，三种模型的正确率既没有随礼貌程度提高而持续上升，也没有因语气更直接而呈现稳定优势。

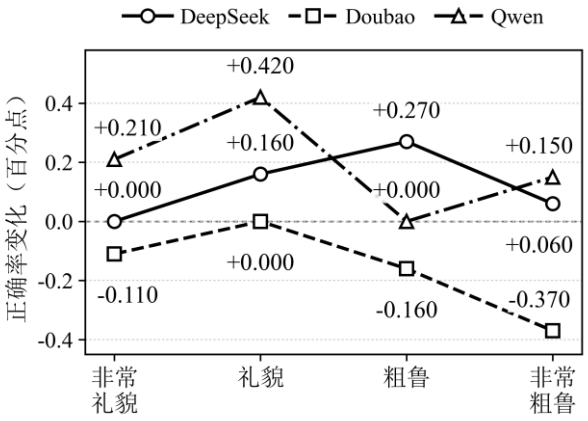

图 2　准确率结果：三种模型相对中性语气的总体变化

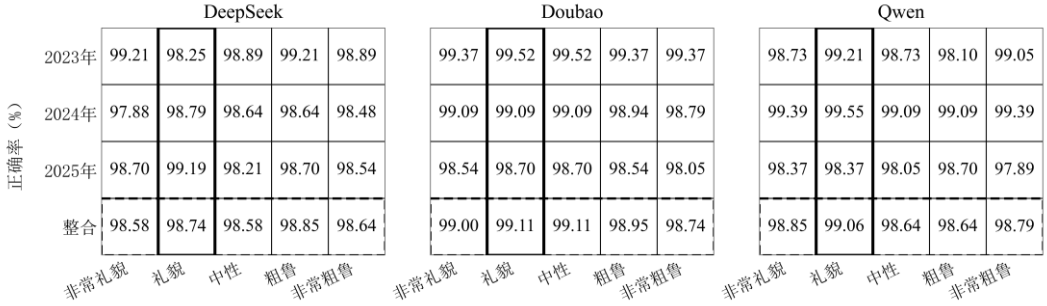

图 3　准确率结果：分模型、分年份与整合口径下的比较

图 3 将上述差异展开到不同模型、不同年份和整合口径。其他非中性语气在局部样本中也会出现高点，但方向随模型和年份变化，难以构成稳定规律。仅保留“至少出现过一次错误”的题目后，困难样本中的语气差异有所放大。这意味着，当题目本身区分度较低、模型正确率接近上限时，语气造成的小幅变化容易被高基线掩盖；在更容易出错的题目中，这种差异才较容易显现。

5．2　语气与词元成本

图4展示了三种模型在四种非中性语气下总词元相对中性语气的变化。横坐标依次为非常礼貌、礼貌、粗鲁和非常粗鲁四种语气，纵坐标为总词元相对中性语气的百分比变化；0%表示与中性语气相同，正值表示总词元增加，负值表示总词元减少。

三种模型的成本变化方向并不一致。DeepSeek在粗鲁语气下的相对总词元为1.1283，即总词元较中性语气增加12.8%；Doubao在四种非中性语气下均高于中性语气，其中非常礼貌语气的增幅最大；Qwen则在非常礼貌和非常粗鲁语气下低于中性语气。语气并不是单纯增加或减少词元，而是会在不同模型中触发不

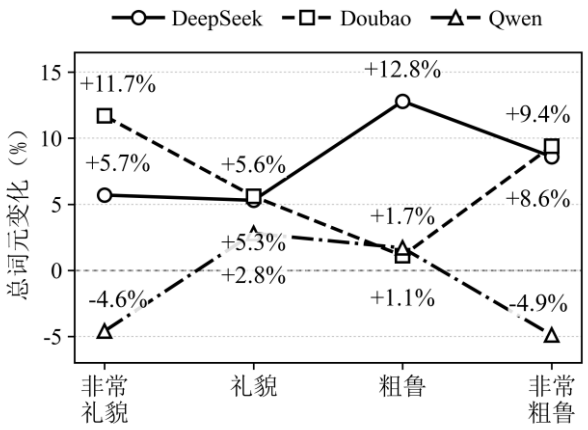

图 4　计算成本结果：总词元相对中性语气的变化

同的资源分配方式。这些成本变化也没有稳定转化为正确率收益。提示词教学若只比较最终答案，容易遗漏这一层差异；课堂实验还应记录词元用量，并检查增加的计算成本究竟用于有效分析，还是仅仅使回答变长。

5．3　语气与回答结构

所有语气条件下，显性社交表达所占比例都接近0。这里统计的是回答文本中可以直接观察到的结构，

不包括单独存储的推理文本。因此，本节讨论的重点不是模型是否回复了礼貌用语，而是语气如何改变正式推理、选项分析、检查总结和题干复述等片段的篇幅。

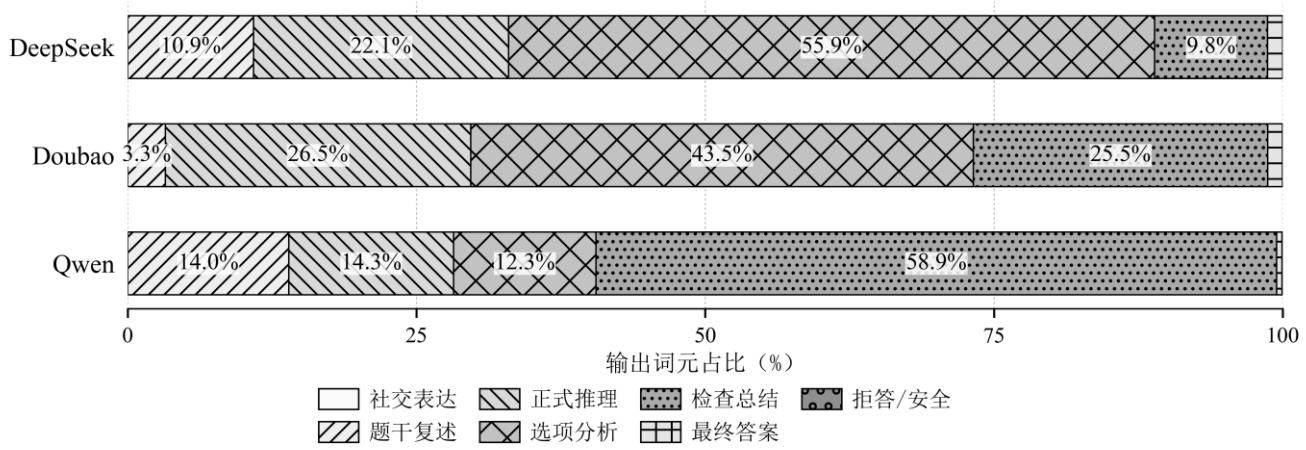

图 5　回答结构结果：三模型默认回答结构的总体占比

图5先给出了三模型的整体分布。可以看到，三模型原有的回答结构并不相同：DeepSeek主要把篇幅放在选项分析上，Doubao在正式推理、选项分析和检查总结之间分配得较为均衡，Qwen则更集中于检查总结。语气并未把它们改造成同一种回答模板，而是在各自原有结构上重新分配篇幅。因而，评价语气效应时不能只比较某一结构的绝对占比，还要考虑模型本身的结构基线。

5．4　教学启示与适用边界

三类结果放在一起后，可以得到一条较清楚的结论：语气带来的差异首先出现在词元投入和回答结构上，只有在部分困难样本中才进一步表现为有限的正确率变化。更准确地说，语气对作答过程的影响比对最终结果的影响更稳定。

高考试题题面统一、答案唯一且便于核验，适合作为提示词课程中的控制变量实验材料。教师可以固定题目内容、模型参数和输出格式，只改变提示词语气，再比较答案、词元用量和回答结构。对于中文客观题或标准化练习，“适度礼貌、任务明确、输出约束清楚”可以作为实验的初始表达，但它不是必须遵守的固定模板。

本文实验使用的样本来自2023—2025年浙江卷纯文本选择题，文本风格正式，答案可以稳定核验；三种模型在成本增减和结构迁移方向上也并不一致。开放问答、写作、多轮对话、数学证明、代码生成以及口语化或含糊指令具有不同的评价标准和交互过程。综合量化评估数据，本文建议采用“适度礼貌”的语气同国产大模型进行交互。

6　结束语

本文以提示词语气为研究对象，将语气划分为五档，重点考察其对DeepSeek、Doubao和Qwen客观题作答的影响。整体来看，在本文样本中，礼貌语气下的正确率不低于中性语气，且相较粗鲁语气表现得更为稳定。研究还发现，国产大模型的词元成本和回答结构对语气变化的响应较为明显，提示词语气更容易改变模型的资源投入和回答组织方式。本文并不主张用户一概采用更礼貌的提示词，而是建议把语气与任务约束、输出结构和词元成本放在同一实验中考察。由于本文只采用高考试题，结论仍有一定局限。标准化题目可用于训练学生控制变量、读取过程指标，并判断新增篇幅是否真正服务于问题求解。后续研究可扩展任务类型和模型范围，检验语气效应在开放、模糊和复杂推理任务中的表现。

参 考 文 献

[1] 国家数据局. 专家解读：数据流通交易加速数据价值释放[EB/OL]. 2026-05-01. https://www.nda.gov.cn/sjj/zwgk/zjjd/0501/20260501112124269218805_pc.html.

[2] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: arXiv preprint arXiv:2201.11903 (2022). doi: 10.48550/arXiv.2201.11903. url: https://arxiv.org/abs/2201.11903.

[3] Jingming Zhuo et al. “ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs”. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, pp. 1950 - 1976. doi: 10.18653/v1/2024.findings-emnlp.108. url: https://aclanthology.org/2024. findings-emnlp.108/.

[4] Ziqi Yin et al. “Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance”. In: Proceedings of the Second Workshop on Social Influence in Conversations. 2024, pp. 9 - 35.

[5] Om Dobariya and Akhil Kumar. “Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy”. In: arXiv preprint arXiv:2510.04950 (2025). url: https://arxiv.org/abs/ 2510.04950.

[6] Hanyu Cai et al. “Does Tone Change the Answer? Evaluating Prompt Politeness Effects on Modern LLMs: GPT, Gemini, and LLaMA”. In: arXiv preprint arXiv:2512.12812 (2025). url: https://arxiv.org/abs/2512.12812.

[7] Axton Liu. Prompt Specificity in River-Crossing Logic: A Comparative Study of General-Purpose vs. Reasoning-Optimized LLMs. Zenodo preprint. 2025. doi: 10.5281/zenodo.15092746. url: https://doi.org/10.5281/zenodo.15092746.

[8] Soogand Alavi, Salar Nozari, and Andrea Luangrath. “Cost Transparency of Enterprise AI Adop-tion”. In: arXiv preprint arXiv:2511.11761 (2025). url: https://arxiv.org/abs/2511.11761.

[9] Neil F. Johnson and Frank Yingjie Huo. “Jekyll-and-Hyde Tipping Point in an AI’s Behavior”. In: arXiv preprint arXiv:2504.20980 (2025). url: https://arxiv.org/abs/2504.20980.

[10] Yixuan Weng et al. “Large Language Models are Better Reasoners with Self-Verification”. In: Findings of the Association for Computational Linguistics: EMNLP 2023. 2023, pp. 2550 - 2575. doi: 10 . 18653 / v1 / 2023 . findings - emnlp . 167. url: https : / / aclanthology . org / 2023 . findings-emnlp.167/.

[11] Hunter Lightman et al. “Let’s Verify Step by Step”. In: arXiv preprint arXiv:2305.20050 (2023). doi: 10.48550/arXiv.2305.20050. url: https://arxiv.org/abs/2305.20050.

[12] Shumin Deng et al. “Towards A Unified View of Answer Calibration for Multi-Step Reasoning”. In: Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations. 2024, pp. 25 - 38. url: https://aclanthology.org/2024.nlrse-1.3/.

[13] Jonathan Uesato et al. “Solving Math Word Problems with Process- and Outcome-Based Feed-back”. In: arXiv preprint arXiv:2211.14275 (2022). doi: 10.48550/arXiv.2211.14275. url: https://arxiv.org/abs/2211.14275.

[14] Hyeong Kyu Choi et al. “How Contaminated Is Your Benchmark? Measuring Dataset Leakage in Large Language Models with Kernel Divergence”. In: Proceedings of the 41st International Conference on Machine Learning. 2025.