arXiv:2603.00958v1 [cs.CL] 1 Mar 2026

S-VoCAL：用于推断文学作品中角色说话声音属性的数据集与评估框架

Abigail Berthe-Pardo[1], **GaspardMichel**[1,2],**Elena V.Epure**[2,3],**Christophe Cerisara**[1]1法国 Vandœuvre-lès-Nancy，LORIA2法国巴黎，Deezer Research3瑞士 Idiap Research Institute
abigail.berthe@hotmail.fr,
gmichel@deezer.com,elena.epure@idiap.ch,christophe.cerisara@loria.fr

摘要

随着文本转语音（TTS）系统的近期进步，合成有声读物旁白受到越来越多的关注，其自然度也达到了前所未有的水平。然而，合成旁白系统在模仿虚构角色，以及传达复杂情感或韵律方面仍存在较大差距。提升角色识别的一种可行方向，是为书中每个虚构角色分配恰当的声音。这一步通常需要在篇幅较长的文本中复杂地推断人物属性，例如角色的年龄、性别、出身或身体健康状况；因此，需要专门的基准数据集来评估抽取系统的性能。我们推出了 S-VoCAL（文学作品中的角色说话声音属性），这是首个专用于评估与声音相关的虚构角色属性推断的数据集与评估框架。S-VoCAL 包含基于社会语音学研究确定的 8 项属性，以及从古腾堡计划中收集的 952 组角色—书籍配对。其评估框架针对每项属性的特点进行了设计，并纳入了一种基于近期大语言模型嵌入的新颖相似度指标。我们通过将简单的检索增强生成（RAG）流程应用于人物属性推断任务，展示了 S-VoCAL 的适用性。结果表明，RAG 流程能够可靠地推断年龄、性别等属性，但在出身或身体健康状况等属性上表现欠佳。数据集和评估代码可在 https://github.com/AbigailBerthe/S-VoCAL。**关键词：**数字人文、语料库、评估方法

1. 引言

有声读物近年来在全球范围内日益受到关注：预计到2025年底，美国、欧洲和中国的有声读物行业收入将增长20%（Statista，2025a，b，c）。有声读物成功的原因之一，常被提及的是它能够将文学作品带给通常不阅读或无法阅读的人群，例如有阅读障碍或视力障碍的人；此外，听书时还能同时进行其他活动（Roy，2024年；Creamer，2025年）。与此同时，文本转语音（TTS）系统合成自然且富有表现力语音的能力不断提升，激发了人们对有声读物合成朗读的兴趣（Walsh 等，2023；Park 等，2025年），使消费者能够聆听人工旁白员为自己朗读喜爱的电子书，也让作者能够以低于传统有声读物制作的成本，自主推广作品。

表演性有声书旁白的一个关键方面，是旁白员能够模仿虚构角色（Browm，2021；Goodwin，2024年）。例如，旁白员通常会在朗读男性角色的话语或叙事部分时采用较低的音高（Pethe et al.，2025年）。对于听众而言，旁白的这一方面对于增强角色认同和阅读理解至关重要（Van Krieken et al.，2017；Wolters et al.，2022），并且已被确定为当前合成有声书的一项关键局限（RoderoandLucas，2023）。此外，现代文本转语音系统在合成虚构人物话语时通常缺乏表现力（Michelet al.，2025年），因此，提升人工朗读系统中的角色模仿能力，成为推动有声书旁白合成取得更广泛进展的核心目标。

可信声音与虚构角色描述之间的对齐开辟了一个颇具前景的方向，该方向近期已在文本转语音社区中得到探索（Xin 等人，2022；Park 等人，2025年）。然而，要找到这些适合角色的声音，需要复杂的自然语言处理（自然语言处理）步骤，以抽取每个角色的声音特征属性，例如年龄、性别、出生地区或身体状况。这些信息在书中未必总是被明确陈述，也可能散布在相距较远的叙事段落中，从而使此类属性的精确抽取颇具挑战（Iosif 和 Mishra，2014；Papoudakis 等人，2024年）。为评估自然语言处理系统准确抽取角色信息的能力，以往的数据集工作大多提出了取自文学分析网站的自然语言一般性角色描述（Yuan 等人，2024年；Gurung 和 Lapata，2024年；Papoudakis 等人，2024年）。此类描述通常以结构化档案或复杂的文学分析为中心，涉及各种与角色声音不相关的辅助信息，或者可能严重遗漏重要的

与声音相关的属性。

因此，我们提出 S-VoCAL（文学作品中的角色说话声音属性），这是一个专门用于抽取与说话声音表现相关的虚构角色属性的数据集与评估框架。S-VoCAL 包含 8 项基于语音学研究的属性，以及 952 个角色-书籍对，分布于从古腾堡计划收集的 192 部不同小说中[1]。

首先通过将角色与其维基数据[2] 条目进行匹配来抽取属性，其中年龄属性则通过人工标注加以补充。鉴于各项属性各具特点且分布不均衡（例如，性别是分类属性；身体健康是简短文本；类型属性大多包含人类），我们设计了专门的评估框架，其中包括加权指标，以及一种基于 Qwen-3 8b 嵌入的新语义相似度指标（Zhang 等人，2025年），并证明该指标与人工判断高度相关。

沿用Papoudakis 等人（2024年）的方法，我们评估了一种简单的检索增强生成（RAG）方法，用于抽取 S-VoCAL 中的角色属性。该方法包括两个步骤：基于指令式编码器生成的嵌入进行段落检索，以及由大语言模型（LLM）执行的属性推断。我们发现，所提出的 RAG 方法通常能够很好地抽取年龄、性别或口语等封闭类属性，但在抽取身体健康和职业等开放类属性时表现不佳。年龄、性别或口语等封闭类属性，但在抽取身体健康和职业。

总而言之，我们的贡献如下：

- 我们提出了 S-VocAL，这是一个用于推断小说中说话者声音角色属性的数据集和评估框架。该数据集已公开，涵盖 192 本不同的书籍中的 952 组人物—书籍配对，并提供 8 种属性。
- 我们设计了一种评估框架，能够考虑每种属性独特且不平衡的特性。
- 我们评估了用于属性推断任务的 RAG 方案，结果表明，该方案在封闭类属性上表现良好，但在推断开放类属性时通常会遇到困难。

2. 相关研究

2.1. 声音特征

声音可以通过音段特征与超音段特征的组合来描述（Nolan，1987；Di Cristo，2013年）。后者对应于音高、响度、节奏等声学特征，这些可以通过文学文本中的各种词汇项目来呈现（Parret，2002）。此外，声音还会传递可用于推断说话者特征的指示性线索（Laver与Trudgill，1979；Abercrombie，1967；Frühholz与Belin，2018年）。在已得到充分研究的特征中，我们发现说话者的性别与年龄（Nolan，1987；Hollien与Shipp，1972a；Sorokowski etal.，2024年）、身体健康状况（Canter，1963b）或出生地区与社会阶层（Labov，1973）。

这些先前的研究使我们能够确定 8 个与语音角色特征相关、且可能从小说中的人物描述中识别出来的显著声学特征。我们将在第 4 节介绍这些属性。

2.2. 人物属性数据集

从叙事文本中提取可靠的角色描述是一项复杂任务，因为相关信息可能是隐含的、会随时间变化，或分散在多个章节中。先前的研究针对角色的特定方面展开，例如社会角色（Stammbach 等人，2022；Gervás 和 Mén- dez，2024年）、潜在人格（Bamman 等人，2014），以及较新的自然语言描述（Brahman 等人，2021；Yuan 等人，2024年；Gurung 和 Lapata，2024年；Papoudakis 等人，2024年）。此类描述可能是对角色的概括总结、文学分析，或围绕特定维度展开（人格、事件等）。与此不同，我们定义了一组与语音表达相关的离散人物属性，并旨在准确提取这些属性，而非高层次的人物描述。

一些研究还致力于提取离散的静态属性集合，例如性别或年龄（Iosif 和 Mishra，2014；Jaipersaud 等人，2024年；Baruah 和 Narayanan，2024年），或动态属性，例如情绪弧线（Vishnubhotla 等人，2024年）。与 S-VoCAL 类似，Iosif 和 Mishra （2014）针对故事讲述应用中的语音相关属性，提出了一个自动提取这些属性的流程。S-VoCAL 在此基础上进行了扩展，研究对象从儿童故事转向小说，并涵盖了源自社会语音学研究的更广泛属性集合。

2.3. 评估属性推断

鉴于年龄或身体健康等属性具有独特且分布不均的特点，评估系统推断角色属性的能力尤其具有挑战性。当属性以自然语言描述时，例如在Papoudakis 等人(2024年的研究中，评估会更具挑战性，因为这需要额外推理，以确保预测描述与真实标注描述在事实上相符。因此，评估

[1] https://www.gutenberg.org/ [2] https://www.wikidata.org/

属性	语音相关性	维基数据属性	Type	覆盖范围	是否为新项
Name		标签	Text	100%	✗
别名		又名、名，本名，昵称，化名	List	73.5%	✗
Age*	音高范围、共振峰频率、语速	相对年龄	分类的（4）	4.1% | 37.7%	✗
性别	音高范围，共鸣	生理性别或社会性别	分类的（2）	96.5%	✗
出身	口音、语音特征，韵律	国籍国、出生地、原籍国	List	58.8%	✓
居住地	口音、言语模式	居住地，工作地点，国家，居住地点	List	26.5%	✓
职业特征性言语	风格	职业，角色类型，担任的职位、贵族头衔、军衔/警衔、船员身份	List	57.7%	✓
口语语言	发音、语调	所讲语言、母语语言	List	49.2%	✓
身体健康	具体的声音特质（例如，气声）	医疗状况	Text	4.1%	✓
Type	人类/非人类声音	实例	Text	100%	✓

表1：S-VoCAL 中的属性概览及其映射的维基数据属性。年龄* 表示为根据语音特征推断出的相对类别，覆盖范围报告为（初始 |最终，经过人工标注）。是否为新项列显示了相较于（Iosif 和 Mishra，2014）的原创性。

通常会使用基于事实的指标来比较预测描述与参考描述，这些指标依据大语言模型（LLM）的判断得出，例如 PRISMA（Mahon 和 Lapata，2024年）或一致性得分（Yuan 等人，2024年）。对于离散属性集，评估方法还包括大语言模型判断（Jaipersaud 等人，2024年）、人类评估（Baruah 和 Narayanan，2024年）以及 F1 分数等标准指标（Iosif 和 Mishra，2014）。

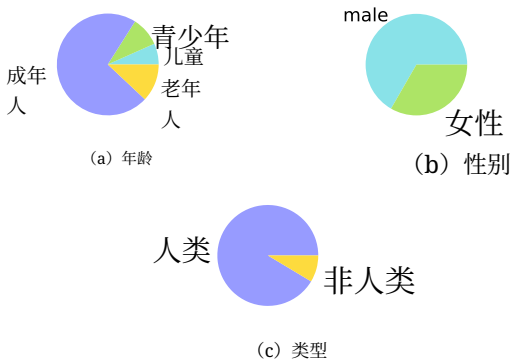
鉴于所选属性的独特性质，我们还采用标准的严格 F1 指标和软F1指标，分别评估封闭类别属性和半封闭类属性（即具有有限可能取值的属性）。对于开放类别的文本属性，我们利用基于 Qwen-3 8b 的指令感知大语言模型嵌入（Zhangetal.，2025）计算预测答案与真实属性之间的语义相似度，方法类似于 BERTScore（Zhangetal.，2019）。指令感知嵌入模型使我们能够得出针对特定属性的表示，从而适应各属性的特性。我们发现，所提出的相似度指标比 BERTScore 更符合人类判断。

3. 属性推断任务

我们定义了一个与语音身份相关的属性分类体系，这些属性可从文本中推断，并可用于为虚构角色设计通用的合成语音。我们的目标是角色身份相关的稳定特征，而非低层次的声学属性。声音携带指示性信息，使听者能够推断出相对稳定的特征，例如年龄、性别、身体状况或地区来源（Abercrombie，1967；Laver andTrudgill，1991）。根据Parret (2002)，这些线索既反映了与生理特征相关的内在层面，也反映了受社会和情境因素影响的外在层面。在合成图书朗读中，这两个维度对于根据虚构角色的人设调整生成语音的声学参数（如音高范围或韵律）都至关重要。

人的特征与其嗓音声学实现之间的联系，已在语音学和社会语音学研究中得到广泛记载。例如，音高范围和基频会随年龄和性别而变化：男性说话者的音高范围低于女性，而女性的音高范围又低于儿童（Cruttenden，1986年），平均音高范围通常会随年龄增长而降低（Hollien andShipp，1972b）。健康状况也会影响嗓音音质和音高稳定性（Canter，1963a），而地理出身被证明与音段实现（即单个音的发音）及口音相关（Labov，1973）。

我们确定了八种可以从叙事文本中推断出的声音相关属性：年龄、性别、籍-



（a）年龄 （b）性别 （c）类型

图 1：由于 S-VoCAL 的属性来自各种古腾堡书籍，因此分布并不均衡，尤其是类型，其中大多是人类；以及年龄，其中大多是成年人。

贯、居住地、口语语言、职业、身体健康以及类型。这些属性对声音特征的具体影响详见表1。

属性推断任务旨在从长篇叙事文本中提取这些与声音相关的属性：给定一个角色和一部小说，我们希望推断出可能影响该角色声音听起来如何的特征。实际操作中，这意味着仅根据书籍内容，预测每种与声音相关的属性的一个或多个取值。这项任务尤其具有挑战性，因为文学文本很少明确陈述此类信息。相关线索往往隐含其中，并分散在相距甚远的段落中，因此需要对长上下文进行推理，才能间接推断某些属性。对超长上下文进行准确推理通常是近期自然语言处理模型难以应对的领域，这为属性推断任务增添了另一层复杂性（Hsieh 等人，2024年；Wang 等人，2024a年）

据我们所知，尚无现有数据集为说话声音属性提供明确标签，也没有用于评估属性推断的通用评估框架。因此，我们提出了 S-VoCAL 数据集，专为属性推断任务而设计，以下将进一步介绍。

4. S-VoCAL 数据集

S-VoCAL（文学作品中的角色语音属性）聚焦于表1中定义的 8 项与声音相关的属性，涵盖 192 部小说中的 952 个角色。这些书籍是 1940年以前出版、属于英语文学或翻译世界文学作品，并且可在古腾堡计划上获取全文的作品[3]。最终选定的作品均满足以下条件：（1）在古腾堡计划上有英语全文版本；（2）在维基数据中有对应条目[4]，；以及（3）在维基数据中至少收录了一个角色（主要或次要角色），且该角色具有用于构建我们属性的目标属性之一（见表 1）。图 1展示了角色在主要分类属性（年龄、性别和类型）上的分布，其中“成年人”“男性”和“人类”角色在语料库中占绝大多数。

4.1. 数据收集

以往研究大多依赖文学摘要网站或粉丝维基中的角色描述（Papoudakis 等人，2024年；Gurung 和 Lapata，2024年；Yuan 等人，2024年）。然而，这些来源在不同书籍和角色的信息详细程度上差异很大，若要利用这些来源，就需要额外构建一套抽取流程，并对其进行验证。因此，我们选择维基数据作为角色信息的主要来源。维基数据是一个协作式知识图谱，提供关于实体的结构化属性，包括由人类贡献者填写的文学作品和虚构角色信息。其结构化格式和频繁的人类编辑活动，使其尤其适合将现有属性与我们的目标属性进行对应。

我们根据维基数据中现有角色的属性整理了数据集。对于每个角色，这些属性各不相同，而且大多无法与我们的目标属性直接对应。我们手动挑选了与目标属性在语义上非常接近的属性（同义词或范围更窄的子集），并将其归入相应的预定义属性。在某些情况下，我们将多个互补值合并到同一属性中。例如，所讲语言汇总了维基数据中的多个属性，例如所讲语言、母语、名称所用语言和书写语言；而属性性别则直接映射到单个维基数据属性生理性别或社会性别。表1详细列出了各属性与维基数据属性之间的映射关系。

收集完成后，我们对数据进行了规范化处理，包括删除多标签属性中的重复项、将相似值转换为标准形式，以及手动整理仍以维基数据编码形式呈现的属性。

如表1所示，不同属性的角色覆盖范围差异很大。具体而言，年龄属性对我们的任务至关重要，但最初只有 40 个实例填写了该属性。因此，我们对部分角色进行人工标注，为其补充年龄值。

4.2. 年龄标注

维基数据中的年龄信息通常以数值形式提供。然而，要塑造角色的声音，并不需要知道其确切年龄，

[3]https://www.gutenberg.org/[4]https://www.wikidata.org

年龄类别	F1 分数	实例
儿童	0.78	16
青少年	0.44	10
成年人	0.90	242
老年人	0.71	25

表 2：人工标注者对年龄 属性的各类别 F1 分数。成年人是最具代表性的类别，其一致性较高；而青少年 的一致性则仅属中等。

而且，通常只有估算出年龄范围，才能提供足够精确的信息。因此，我们将年龄属性转换为分类属性，分为四类：儿童、青少年、成年人和老年人。对于年龄以数值形式提供的 40 个角色，我们自动将其映射到相应类别。随后，我们又手动标注了 319 个年龄实例，共计 359 个标注案例。我们根据提供角色分析的网站（例如 GradeSaver[5]，LitCharts[6]）推断年龄类别。我们设计了[7] 指南，优先采纳明确提及的年龄，其次参考相对年龄线索、清晰的指称（如“儿童”）以及人生阶段指标。对于无法确定的案例，则留空。

共有 293 个角色接受了双重标注，我们使用标准版和软变体的科恩 κ 系数与 F1 分数来衡量标注员间的一致性。与其他封闭类属性不同，年龄具有线性结构，从“儿童”到“老年人”自然递进。因此，相邻类别之间的差异（例如将“青少年”标为“成年人”）比相距较远的类别差异（例如将“儿童”标为“老年人”）轻微得多，而且对于某些角色而言，这可能是难以在两个相邻类别之间作出选择所致。基于这一原因，我们还报告了两项指标的软版本：当两项标注仅相差一个相邻年龄类别时，给予 0.8 的权重，计为部分得分。

我们报告 Cohen’s κ 系数的标准形式为 0.56，采用软变体计算时则提升至 $\kappa = 0.74$，同时报告软 F1 分数为 0.86。这些观察结果表明标注者之间具有中等程度的一致性，也凸显出大多数分歧发生在相邻类别之间。表 2 展示了 293 个双重标注实例的各类别 F1 分数，结果表明“青少年”是分歧最多的年龄类别。标注者随后进行了裁定，以解决剩余分歧并完善指南。经过这一流程后，359 个角色已填写年龄属性。数据集中各年龄类别的分布见图 1。

[5]https://www.gradesaver.com/[6] https://www.litcharts.com/[7]附录 A

5. S-VoCAL 评估框架

我们提出了一套专门的评估框架，针对不同属性类型和数据格式（封闭、半封闭和开放）量身定制。封闭类和半封闭类属性采用基于 F1 的指标进行评估，而开放属性则采用基于 Qwen-3 嵌入的相似度指标进行评估。该指标类似于 BERTScore，并通过与人工判断的相关性进行验证。我们还利用这些判断将相似度值映射到人类可理解的尺度，以便更好地理解模型在开放属性上的表现。

5.1. 封闭类和半封闭类评估

封闭类属性（年龄、性别和类型（人类/非人类））的取值范围较小，因此可以直接使用标准分类指标。由于 S-VoCAL 中大多数角色的类型是“人类”（91%），我们对这一属性进行两阶段评估。第一阶段评估系统区分人类和非人类角色的准确度，第二阶段则考量非人类角色之间的语义相似性。类型（人类/非人类）、年龄和性别均采用加权 F1 进行评估，以应对各类别中的标签不平衡问题。

与年龄的人工标注方法类似，我们提出一种软 F1 分数，对相邻类别中的错误给予较轻的惩罚。自定义软 F1 分数使用权重矩阵来编码年龄组之间的线性距离：精确匹配为 1.0，相邻类别为 0.8，其余情况为 0.0。

唯一的半封闭类属性口语语言的取值范围更广，但仍限定于语言名称集合，因此与开放属性相比，语义歧义较小。因此，可以使用基于取值精确匹配的指标（如 F1 分数）对其进行评估。我们将口语语言作为多标签设置进行评估，先按类别进行二值化，再计算微平均 F1；这种方法适合评估系统推断每个角色所讲语言集合（可能包含多种语言）的能力。

5.2. 开放类评估

开放类属性（出身、居住地、职业、身体健康、类型（非人类））可以有多种表达方式，同时含义却非常

属性	预测	标准参考答案	人工评分	解释尺度
出身	美国	法国	0.0	预测完全错误
职业	作家	作者，护士	0.5	部分正确，但缺少一些职业
身体健康	言语障碍	口吃	0.66	不够精确，但正确
Type	小猫	Cat	1.0	正确或同义词

表3：标注者用于评估各系统对开放类属性的预测答案的解释尺度示例。

相近，例如，对于居住地，可以说“英格兰”或“英国”；对于职业，可以说“贵族”或“伯爵”。BERTScore（Zhang 等人，2019年）已广泛用于生成系统的评估，也可以从理论上应用于我们的评估框架。然而，在 S-VoCAL 开放类属性的情境下应用 BERTScore，可能会面临以下局限。首先，我们的标准参考答案和系统预测通常是词语或实体列表，或是短语，缺乏上下文依赖关系，因而限制了基于 BERT 的表征能力（Ethayarajh，2019）。其次，有报告指出，BERTScore 无法忠实评估摘要的文本质量（Fabbri 等人，2021），这促使研究界转向基于大语言模型的自动评估指标（Min 等人，2023）。

基于 Qwen3 的指标 为了解决这些局限性，我们提出使用基于 Qwen3-8b 嵌入[8], 的属性感知表示来评估标准答案与预测答案之间的语义相似度。Qwen3-8b 嵌入是一种最新的感知指令的嵌入模型，能够根据指令计算文本表示。我们利用其对指令的敏感性，为每个属性设计简短的提示，从而得到以属性为条件的标准参考答案和预测输出表示。随后，我们直接计算以属性为条件的标准答案表示与预测答案表示之间的余弦相似度。

验证 我们通过计算余弦相似度分数与人工判断之间的斯皮尔曼 ρ 相关性，验证所提出指标的质量。为获得这些判断，三位标注员按照 0 到 1 的离散量表，对预测出的开放类属性子集进行评估，其中 0 表示预测完全错误，1 表示预测正确，中间分数则表示部分正确。为确保标注一致，我们向每位标注员提供了各个分值对应的可解释判断，并辅以示例（部分示例见表3，指南见附录 B）。

[8]https://huggingface.co/Qwen/Qwen3-Embedding- 8B

属性	α	ρ-Qwen3	ρ-BERT	n
出身	0.95	0.85	0.44	66
居住地	0.93	0.85	0.31 ($p = 0.013$)	65
身体健康	0.74	0.52	-0.10 ($p = 0.572$)	37
职业	0.75	0.66	0.32	121
类型（非人类）	0.92	0.86	0.63	57

表 4：标注者间一致性（α）以及人工评分与自动指标之间的斯皮尔曼相关性（ρ）。除非另有说明，所有相关性在 $p < 0.001$ 水平上均显著缩写：身体健康 = 身体健康；类型（非人类） =类型（非人类）。

我们报告标注者间一致性，采用区间克里彭多夫 α（Krippendorff，2011）以及斯皮尔曼 ρ 相关系数，以衡量人工判断与基于 Qwen3 的指标和 BERTScore 之间的相关性，见表4。人工标注在所有属性上均表现出较高一致性，克里彭多夫α系数范围为 0.74 至 0.95。此外，所提出的指标在所有属性上与人工判断的相关性均持续且显著高于 BERTScore，斯皮尔曼ρ 范围为 0.52 至 0.85，而 BERTScore 在身体健康上的相关性为零，最大值为 $\rho = 0.63$。有趣的是，基于 Qwen3 的指标与人工判断之间的相关性在身体健康和职业属性上较低，而这些属性的标注者间一致性也较低，这可能表明比较这些属性的标准答案与预测值更具挑战性。

人类对齐分数 虽然上述相关性分析表明，与 BERTScore 相比，我们提出的指标在评估开放属性时更加可靠，但原始余弦相似度分数缺乏直接的可解释性。因此，我们进一步利用人工判断，通过拟合一条保序回归线，将 Qwen3 的相似度值映射到每个属性对应的人类评估尺度上，从而计算人类对齐分数（HAS）。这一过程会自动将原始相似度分数转换为易于理解的值，使模型预测更清晰易懂。需要注意的是，这种方法只能对



图 2：检索增强生成流程概览。第一步为角色提取属性相关摘录。随后，将这些段落连同属性提示一起输入大语言模型。

模型预测作部分解释，并且会受到保序回归中混杂因素的影响。不过，我们通过实证发现，HAS 通常与人工判断高度一致，且比余弦相似度分数更容易理解，因此建议在 S-VoCAL 的评估框架中使用 HAS。

6. 使用 RAG 进行属性推断

遵循先前的研究（Papoudakis 等人，2024年），我们使用检索增强生成（RAG）对 S-VoCAL 进行属性推断。给定一部小说及角色的标识信息（名称、别名或昵称），检索增强生成流程会从相关书籍摘录中推断角色属性。检索增强生成（RAG）尤其适用于属性推断场景：与角色属性相关的信息可能散见于书中的不同部分；在最糟糕的情况下，整部叙事中可能只隐晦地提及一次。通过检索并处理与属性相关的书籍段落，我们解决了一次性处理整本书时出现的“大海捞针”问题。检索增强生成流程如图 2所示，由三个主要组件构成，下面的小节将对此进行介绍。

6.1. 检索角色摘录

检索组件根据语义相似度，选取提及该角色的最相关段落。首先，我们利用角色的名称和别名，通过正则表达式定位小说中所有对该角色的具名提及，然后在每次提及前后提取 200 词的窗口。我们通过实验将这一数量设定为 200，因为它在捕捉相关信息与将上下文大小控制在模型限制范围内之间取得了良好平衡。随后，我们计算使用 E5-large 生成段落嵌入[9]（Wang 等人，2024b），选择该模型是因为其架构轻量，并且支持基于查询的句子嵌入；此外，它近期已应用于文学分析，并在文学语境中展现出良好效果（Hatzel 和 Biemann，2024年）。为提取特定于属性的相关摘录，我们执行以下步骤：

- **说明：**我们定义了一个通用指令提示，适用于所有属性查询
- **属性查询：**针对每个目标属性，都会创建一个特定查询（例如，对于年龄：“该角色是儿童、青少年、成年人还是老年人？”）。所有查询见附录C。
- **评分：**我们对所有查询和段落表示进行编码，然后计算二者之间的余弦相似度。
- **筛选：**我们为每个属性选取最相似的十个段落；当请求多个属性时，会去除各前10列表中的重复项。该列表将用作基于大语言模型的属性推断步骤的输入。

6.2. 属性推断

利用检索步骤得到的与属性相关的摘录，我们提示大语言模型推断所请求的人物属性。针对每个角色构建单一提示，将所有请求的属性合并到一次查询中。在本研究中，我们使用Qwen3-8B（Yang等人，2025年）；以及Phi-4 14B（Abdin等人，2024年）作为大语言模型骨干。两种模型均被量化为4比特（Q4_K_M），以优化

[9]https://huggingface.co/intfloat/multilingual-e5-large-instruct

		Qwen3-8b	Phi4-14b 基线	
Weighted F1	性别	0.993	**0.997**	0.561
	Age	0.783	**0.800**	0.660
	类型（人类/非人类）	0.969	**0.971**	0.910
Soft F1	Age	0.927	**0.941**	0.908
Micro F1	口语语言	0.712	**0.749**	0.576

表5：封闭类别属性和半封闭类属性的结果及其专用指标。类型（人类/非人类）对应于“人类”与“非人类”类型的二元分类。

内存使用量和推理速度，并且我们保留了默认生成参数。

我们观察到，模型输出往往包含噪声且格式不完整（例如，含有推理痕迹、Markdown 瑕疵或格式错误的 JSON 结构）。因此，在将预测结果转换为结构化角色属性之前，我们会进行后处理以清理预测结果（详情见附录 D ）。

7. 结果

我们通过所提出的评估框架，对基于 RAG 的流水线在 S-VoCAL 上的表现进行了评估，并报告了结果。

7.1. 封闭类与半封闭类结果

对于封闭类和半封闭类属性，我们将大语言模型的表现与多数值基线进行比较；该基线始终为 S-VoCAL 中的每项属性预测出现频率最高的值。该基线为基于 RAG 的流水线所得结果提供了最低限度的参照，便于在上下文中理解这些结果。

结果见表5。对于封闭类属性，两个基于 RAG 的流水线表现都不错，且显著优于多数值基线。以 Phi-4 14B 作为大语言模型骨干时，得分略高于使用 Qwen3-8B 时的得分。两个系统在性别和 类型上均取得了很高的加权 F1 分数，而年龄方面的结果则略低。使用软F1评估年龄属性时，结果有所提升，这表明大多数错误都发生在相邻类别之间；人工标注中也报告了这一现象。

对于半封闭属性“口语语言”，两种基于 RAG 的流水线都取得了相当可观的

	mHAS ↑	
属性	Qwen3-8b	Phi4-14b
职业	**0.522**	0.486
居住地	0.498	**0.502**
出身	0.420	**0.422**
类型（非人类）	**0.785**	0.781
身体健康	0.141	**0.165**

表6：开放属性的人类对齐平均分（mHAS）。对于大多数属性，0.5 表示预测部分正确，但信息不完整。类型（非人类） =类型（非人类）。

较基线有所提升，分数表明大多数语言都得到了正确推断。不过，与最高分之间仍存在显著差距，这表明推断出的口语语言集合中偶尔会有遗漏或错误。

7.2. 开放类别结果

我们在表 6 中报告了人类对齐平均分（mHAS），该分数根据第5.2节所介绍的相似度度量计算得出。我们选择报告 mHAS 而非原始相似度分数，以便于理解；如需了解原始相似度分数，请参阅附录 E 。mHAS 的取值范围为 0 到 1，其中 0.5 通常表示答案部分有效。各属性取值的详细解读请参阅附录B。

类型 的 mHAS 最高（≈ 0.78），对应于大多数预测正确，但有些情况下缺乏精确度（例如，预测为“鸟”而非“乌鸦”）。职业 （≈ 0.48–0.52）表明推断部分正确，但角色的一些职业仍未出现在预测答案中。居住地（≈ 0.5）对应于接近预期值但粒度更粗的预测，例如国家正确但地区错误；也可能是预测部分正确，但遗漏了一些应预测的地点。出身 （≈ 0.42）介于高层级正确（大洲正确但国家错误）和部分重合（一些地点正确，但也有遗漏）之间。身体健康 （≈ 0.15）低于第一个解读层级（0.33 = 对健康状况不佳与良好作出模糊但正确的区分），这意味着系统往往无法判断角色是否存在健康问题。这些结果表明，尽管类型 能够被很好地识别出来，职业、居住地和出身 的预测往往只是部分正确，而身体健康 仍然具有挑战性

在两条大语言模型流程之间，检索开放类属性并无明显胜者：我们仅观察到各属性之间存在轻微波动，唯有职业似乎由 Qwen3-8b 推断得略好一些。有趣的是，我们观察到

Qwen3 与 Phi4 的表现之间并无明显的线性关系：若某一模型在特定实例上表现良好，并不必然意味着另一模型在该实例上同样表现良好，这表明两个模型具有相互独立的预测行为。尽管这种独立性背后的原因尚不明确，我们推测这两个模型可能呈现出不同的书籍记忆模式（Chang 等人，2023），这可能导致某一模型在另一模型未曾记忆的某本书的实例上表现良好，从而表明其在该书上的表现较差。我们在附录 F中进一步详述这一点。

8. 结论

本文介绍了 S-VoCAL，这是一个数据集及评估框架，旨在评估能否从小说中推断出与角色声音设计相关的人物属性。S-VoCAL 涵盖了基于社会语音学研究的一系列广泛属性，并收录了 192 本书和 952个角色。鉴于属性的异质性，我们设计了一个全面的评估框架，其中包括适用于封闭类和半封闭类情形的标准指标，以及一种基于近期大语言模型、用于开放类属性的新型语义相似度指标。因此，S-VoCAL 为评估旨在从小说中推断与声音相关的人物属性的系统提供了标准化基准。我们通过应用一个简单的 RAG 流程，展示了 S-VoCAL 的适用性；该流程能够从整部小说中推断指定角色的声音相关属性。结果表明，该流程能够可靠地处理封闭类属性（例如，年龄、性别、类型），但在处理开放属性（如 身体健康或出身）时仍有局限。S-VoCAL 有助于创建沉浸式的合成图书旁白，为每个角色分配与其身份相符的声音。

9. 局限性

尽管本研究致力于构建一个用于推断虚构角色与声音相关属性的数据集与评估框架，但仍存在若干局限性。

首先，S-VoCAL 仅包含取自古腾堡计划的公开书籍。这些书籍往往是经典文学作品，大多数大语言模型（LLM）很可能都记忆过这些作品。在评估基于大语言模型的属性推断时，这一局限性至关重要，因为它们在 S-VoCAL 上报告的性能并不一定意味着它们在近期出版、预训练期间未见过的书籍上也能表现出色。不过，我们认为，S-VoCAL 的结构化评估框架仍然适用于评估这类近期出版的书籍。另一种可能是，对于预训练期间见过的书籍，大语言模型对其内容的记忆以及由此扩展的知识——这些因素会提高预测属性的准确性——可能在广为人知的角色的忠实声音设计中发挥重要作用。

此外，我们的人工标注仅覆盖了年龄属性的一个子集，因为我们认为它是设计忠实角色声音最关键的属性。然而，S-VoCAL 中的其他属性在很大程度上仍代表性不足，将这一标注扩展到这些属性将非常有助于提高所有属性类型评估的覆盖范围与可靠性。

此外，S-VoCAL 提供了 8 项源自维基数据的静态角色属性。然而，许多所选属性可能会随时间演变，例如年龄、身体健康或所讲语言。我们发现，从任何来源（无论是维基数据还是分析网站）恢复这些动态模式都特别具有挑战性，因此转而选择静态属性，作为评估属性推断系统的最低基准。然而，我们认为属性动态建模对于在整个故事中忠实呈现角色声音至关重要，并希望 S-VoCAL 能成为迈向这一目标的第一步。

最后，除所选属性之外的其他角色特性也可能影响角色的声音。然而，这些特征要么高度依赖语境（例如其情绪），要么取决于读者的诠释判断，因此在结构化资源中的呈现一致性较低。我们认为，S-VoCAL 为通用角色声音设计迈出了第一步；未来还可以通过额外的推断层进行基于语境的进一步呈现，这些将留待后续研究探索。

10. 伦理考量

本文介绍的数据集可能带有结构性偏差，因为古腾堡计划，尤其是其中的英语书籍，主要收录西方文学经典，而维基数据中的可用信息在不同实体之间仍不均衡。这些不平衡可能影响最常出现并因此得到评估的角色类型和属性。

此外，未来关于合成语音生成的研究应通过制定明确的伦理指南和监测实践，确保所定义属性的使用保持负责任，从而避免在无意中将角色特征与说话风格刻板地关联起来。

最后，由于 S-VoCAL 数据集完全基于公开可获取的书籍构建，因此仍然

存在不平衡问题，其中成年男性人类角色占据主导地位。这种分布并不能完全反映文学作品中角色的多样性，未来研究纳入更加多样化的书籍，或许有助于弥补这一局限。

11. 书目参考文献

Marah Abdin、Jyoti Aneja、Harkirat Behl、Sébastien Bubeck、Ronen Eldan、Suriya Gunasekar、Michael Harrison、Russell J. Hewett、Mojan Javaheripi、Piero Kauffmann、James R. Lee、Yin Tat Lee、Yuanzhi Li、Weishung Liu、Caio C. T. Mendes、Anh Nguyen、Eric Price、Gustavo de Rosa、Olli Saarikivi、Adil Salim、Shital Shah、Xin Wang、Rachel Ward、Yue Wu、Dingli Yu、Cyril Zhang 和 Yi Zhang。2024年。Phi-4 技术报告。

David Abercrombie。1967。普通语音学要论。爱丁堡大学出版社，爱丁堡。

David Bamman、Ted Underwood 和 Noah A Smith。2014。文学角色的贝叶斯混合效应模型。载于计算语言学协会第 *52* 届年度会议论文集（第 *1* 卷：长篇论文），第 370–379 页。

Sabyasachee Baruah 和 Shrikanth Narayanan。2024年。使用大语言模型从电影剧本中抽取角色属性。载于 *ICASSP 2024 - 2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 8270–8275 页。

Faeze Brahman、Meng Huang、Oyvind Tafjord、Chao Zhao、Mrinmaya Sachan 和 Snigdha Chaturvedi。2021年。“让你的角色讲述他们的故事”：一个用于以角色为中心的叙事理解的数据集。*arXiv*预印本 *arXiv:2109.05438*。

JamesBrowm。2021年。《讲故事的艺术——画外音如何让有声读物栩栩如生》。https://jamesbrownvoice.com/voice-over-bringing-audiobooks-to-life。

Gerald J. Canter。1963a。帕金森病患者的语音特征。一、强度、音高和时长。《言语与听力障碍期刊》，28:221–229。

Gerald J. Canter。1963b。帕金森病患者的语音特征：I. 强度、音高和时长。言语与听力障碍杂志，28(3)：221–229。

Kent Chang、Mackenzie Cramer、Sandeep Soni 和 David Bamman。2023。开口说话，记忆：对 ChatGPT/GPT-4 已知书籍的考古研究。载于*2023* 年自然语言处理实证方法会议论文集，第 7312–7327 页，新加坡。计算语言学协会。

Ella Creamer。2025年。英国去年有声书收入增长近三分之一。https://shorturl.at/5IRYk。

Alan Cruttenden。1986年。语调。剑桥大学出版社，剑桥。

Albert Di Cristo。2013年。言语韵律。De Boeck Superieur。

Kawin Ethayarajh。2019。上下文化词表示的上下文性有多强？比较 BERT、ELMo 和 GPT-2 嵌入的几何结构。收录于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第55–65页，中国香港。计算语言学协会。

Alexander R. Fabbri、Wojciech Kryściński、Bryan McCann、Caiming Xiong、Richard Socher 和 Dragomir Radev。2021。Summeval：重新评估摘要评测。

Sascha Frühholz 和 Pascal Belin。2018年。《牛津声音感知手册》。牛津大学出版社。

Pablo Gervás 和 Gonzalo Méndez。2024年。使用大语言模型标注叙事中的普罗普人物功能。载于*Text2Story@ECIR*，第 137–148 页。

Brittany Goodwin。2024年。如此，你想要讲述有声读物。https://www.brittanygoodwin.com/blog/2019/12/8/so-you-want-to-narrate-audiobooks。

Alexander Gurung 和 Mirella Lapata。2024年。Chiron：长篇叙事中的丰富人物表示。载于计算语言学协会：*EMNLP2024* 论文集，第 8523–8547 页。计算语言学协会。

Hans Ole Hatzel 和 Chris Biemann。2024年。故事嵌入——以叙事为核心的虚构故事表征。载于*2024*年自然语言处理实证方法会议论文集，第5931–5943页，迈阿密，佛罗里达州，美国。计算语言学协会。

Harry Hollien 和 Thomas Shipp。1972a。男性的言语基频与实足年龄。《言语与听力研究期刊》，15(1):155–159。

Harry Hollien和Thomas F. Shipp。1972b。男性说话基频与实际年龄。言语与听力研究期刊，15:155–159。

Cheng-Ping Hsieh、Simeng Sun、Samuel Kriman、Shantanu Acharya、Dima Rekesh、Fei Jia、Yang Zhang和Boris Ginsburg。2024年。Ruler：长上下文语言模型的真实上下文长度是多少?

Elias Iosif和Taniya Mishra。2014。从说话人识别到情感分析：一种分析儿童故事的多步骤系统。载于第三届文学计算语言学研讨会（*CLFL*）论文集，第40–49页，瑞典哥德堡。计算语言学协会。Brandon Jaipersaud、Zining Zhu、Frank Rudzicz和Elliot Creager。2024年。展示而非讲述：利用大语言模型揭示隐性人物刻画。Klaus Krippendorff。2011。计算克里彭多夫α系数的信度。William Labov。1973。社会语言学模式。第4卷。宾夕法尼亚大学出版社。John Laver和Peter Trudgill。1979。言语中的语音与语言标记。言语中的社会标记，1:32。John Laver和Peter Trudgill。1991。言语中的语音与语言标记。载于编辑John Laver编，言语的馈赠，第235–264页。爱丁堡大学出版社，爱丁堡。Louis Mahon和Mirella Lapata。2024年。一种用于电视剧多模态摘要的模块化方法。Gaspard Michel、Elena V. Epure和Christophe Cerisara。2025年。Libriquote：用于表现力零样本语音合成的虚构人物话语语音数据集。Sewon Min、Kalpesh Krishna、Xinxi Lyu、Mike Lewis、Wen-tau Yih、Pang Koh、Mohit Iyyer、Luke Zettlemoyer和Hannaneh Hajishirzi。2023。FActScore：长文本生成中事实精确度的细粒度原子级评估。载于*2023*年自然语言处理实证方法会议论文集，第12076–12100页，新加坡。计算语言学协会。

FrancisNolan.1987. Alan Cruttenden，《语调》。剑桥：剑桥大学出版社，1986年，第 xiv 页 + 214。 语言学期刊，23(2):455–458。Argyrios Papoudakis、Mirella Lapata 和 Frank Keller。2024。BookWorm：一个用于人物描述与分析的数据集。载于 计算语言学协会：*EMNLP 2024* 论文集，第 4471–4500 页，美国佛罗里达州迈阿密。计算语言学协会。Kyeongman Park、Seongho Joo 和 Kyomin Jung。2025。 Multiactor-audiobook：使用多位说话者的面孔与声音进行零样本有声读物生成。Herman Parret。2002。 第 *2* 章：声音的品质，《哲学论点》，第 39–66 页。De Boeck Supérieur，新鲁汶。Charuta Pethe、Bach Pham、Felix D Childress、Yunting Yin 和 Steven Skiena。2025。 有声读物的韵律分析。载于*2025*年第*19*届国际语义计算会议（*ICSC*），第 217–221 页。IEEE。Emma Rodero 和 Ignacio Lucas。2023。 有声读物中的合成语音与真人语音：人类情感亲密效应。 新媒体与社会，25(7):1746–1764。Nilanjana Roy。2024。有声读物算不算阅读？ 金融时报。Piotr Sorokowski 、 Katarzyna Pisanski、Tomasz Frąckowiak、Aleksander Kobylarek 和 Agata Groyecka-Bernard。2024。 基于自由言语与脚本言语对生理性别、身高、体重、吸引力、健康状况和心理特征的语音判断。 心理计量通报与评论，第 1–10 页。Dominik Stammbach、Maria Antoniak 和 Elliott Ash。2022。 英雄、反派与受害者，以及 GPT-3：无需训练数据即可自动抽取角色角色。Statista。2025a。有声读物——中国。https://www.statista.com/outlook/amo/media/books/audiobooks/china。Statista。2025b。有声读物——欧洲。https://www.statista.com/outlook/amo/media/books/audiobooks/europe。Statista。2025c。有声读物——美国。https://www.statista.com/outlook/amo/media/books/audiobooks/united-states。

Kobie Van Krieken、Hans Hoeken 和 José Sanders。2017。唤起并测量对叙事人物的认同——语言线索框架。《心理学前沿》，8:1190。

Krishnapriya Vishnubhotla、Adam Hammond、Graeme Hirst 和 Saif Mohammad。2024年。文学小说的情绪动态。载于计算语言学协会论文集：*ACL 2024*，第2557–2574页，泰国曼谷。计算语言学协会。

Brendan Walsh、Mark Hamilton、Greg Newby、Xi Wang、Serena Ruan、Sheng Zhao、Lei He、Shaofei Zhang、Eric Dettinger、William T. Free-man 和 Markus Weimer。2023年。大规模自动有声书创建。

Chonghua Wang、Haodong Duan、Songyang Zhang、Dahua Lin 和 Kai Chen。2024a年。Ada-leval：使用长度自适应基准评估长上下文大语言模型。

Liang Wang、Nan Yang、Xiaolong Huang、Binx-ing Jiao、Linjun Yang、Daxin Jiang、Rangan Ma-jumder 和Furu Wei。2024b年。通过弱监督对比预训练生成的文本嵌入。

Alissa P Wolters、Young-Suk Grace Kim 和 John William Szura。2022年。《阅读韵律是否与阅读理解相关？一项元分析》。《阅读科学研究》，26(1):1–20。

Detai Xin、Sharath Adavanne、Federico Ang、Ashish Kulkarni、Shinnosuke Takamichi 和 Hiroshi Saruwatari。2022年。利用声学和文本上下文改善有声书文本转语音合成的语音韵律。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chu-jie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、TianyiTang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。2025年。Qwen3技术报告。

Xinfeng Yuan、Siyu Yuan、Yuhan Cui、Tianhe Lin、Xintao Wang、Rui Xu、Jiangjie Chen 和 Deqing Yang。2024年。评估角色理解通过从虚构作品中提取角色档案来评估大语言模型。载于*2024*年自然语言处理实证方法会议论文集，第 8015–8036 页，美国佛罗里达州迈阿密。计算语言学协会。

Tianyi Zhang、VarshaKishore、FelixWu、Kilian Q. Weinberger和Yoav Artzi。2019年。BERTScore：使用BERT评估文本生成。*CoRR*，abs/1904.09675。

Yanzhao Zhang、Mingxin Li、Dingkun Long、Xin Zhang、Huan Lin、Baosong Yang、Pengjun Xie、An Yang、Dayiheng Liu、Junyang Lin、Fei Huang 和 Jingren Zhou。2025年。Qwen3 Embedding：通过基础模型推进文本嵌入和重排序。

附录

A. 角色年龄标注指南

下面介绍提供给人工参与者的指南，用于标注角色年龄。角色分为四个年龄类别：

- **儿童**：0 至 12 岁
- **青少年**：13 至 17 岁
- **成年人**：18 至 59 岁
- **老年人**：60 岁及以上

A.1. 来源

必须在线使用角色的全名和书名查询角色，以便从 Wikipedia 、 SparkNotes 、 GradeSaver 、 LitCharts 等知名网站等网络来源推断其年龄。这类网站通常包含书籍的情节概要、人物描述或分析，可据此推断角色的年龄。为了确保信息可靠，鼓励通过多个来源交叉核查标注。如果多个来源提供的信息相互矛盾，应选择最一致的描述；如果仍无法确定，则将标注留空。

A.2. 一般原则

年龄类别应按照以下优先顺序确定：

1. 文本中明确提及的年龄。2. 相对年龄信息（例如，“比他 18 岁的姐姐年轻”）。3. 直接描述性术语或关键词（例如，“男孩”“老人”“青少年女孩”）。4. 行为、角色或生活状况（例如，在学校就读、有稳定工作、已婚或为人父母）。

A.3. 标注决策步骤

1. **是否明确给出了数值年龄？** → 使用上表为其分配类别。

2. **如果没有，年龄是否是相对于另一个角色给出的？**

   示例：“比他 18 岁的姐姐年纪小”，某人被描述为比一位已标注为老年人的人年长。→推断最可能的类别。

类别	**常用关键词/表达-式**
儿童	婴儿、婴儿、幼儿、男孩、女孩、小的，小孩、男学生、女学生
青少年	青少年、青少年、高中学生，男学生（高中学校），青年男子、青年女子（也可能根据上下文为“成年人”）
成年人	man，女子，丈夫，wife，工人，士兵，家长，教师，商人
老年人	old man，old 女子，年老，祖父、祖母、退休者，体弱者、年迈者

表7：用于推断年龄类别的常见关键词。

3. **如果不是，是否存在明确的描述性术语？** 示例：“儿童”、“女子”、“老人”。→据此分配类别。

4. **如果没有，行为或社会角色能否表明年龄？**

   - 正在上小学 →儿童。
   - 正在上高中 →青少年。
   - 有稳定工作或已婚 →成年人。
   - 被描述为“年老”或“体弱者” →老年人。

5. **如果无法得出合理结论：** 将标注留空。

4. 非人类实体的特殊规则

- 如果年龄有明确说明且符合现实（例如精灵“20岁”），则应用与人类相同的规则。
- 如果年龄不符合现实（例如“200岁的吸血鬼”），则根据角色的行为和外貌分配类别。
- 如果未说明年龄，则应用与人类相同的决策步骤。

5. 常见关键词示例

我们在表7 中列出了年龄类别所使用的常见关键词。

B. 评估模型预测的指南

下面介绍我们提供给人类参与者的指南，用于评估开放类别属性中的模型预测。这些指南包括每个类别的示例，可用于解读我们的平均人类对齐分数。

B.1. 通用评估设置

每个属性定义了三个要素：

- {attribute}_预测答案：模型的预测。
- {attribute}_标准答案：标准参考答案（正确答案）。
- {attribute}_一致性：标注员给出的一致性分数。

仅标注一致性分数。分数从针对每种属性预先设定的一组值中选取。

B.2. 籍贯和居住地

籍贯和居住地作为列表属性，按照以下一致性量表进行评估：

- **0.0**：预测答案完全错误（例如，与问题无关，或大洲错误）。
- **0.25**：预测答案在较高层级上部分相关（例如，大洲正确但国家错误）。
- **0.5**：预测答案在较低层级上部分正确（国家正确，但地区或子区域错误），或者部分正确但遗漏了一些地点。
- **0.75**：预测答案正确，但不如标准答案具体；或者过于具体，无法完全核实。
- **1.0**：预测答案完全正确，与标准答案完全对应，也包括同义词。对于城市，接受彼此非常接近的城市。如果标准答案是一个国家，则接受该国的首都或境内非常大的城市。

示例

- 0.0：标准答案 = ["法国"],预测答案 = ["美国"]
- 0.5：标准答案 = ["法国", "布列塔尼"]，预测答案 =["法国", "马赛"]
- 1.0: 标准答案 = ["法国"]，预测答案 = ["法国", "巴黎"]

B.3. 身体健康

- **0.0**: 预测不相关，或状态判断（好/差）错误。
- **0.33**：预测较为模糊，但正确识别了健康/不健康。
- **0.66**：正确，但不如标准答案具体。
- **1.0**：完全正确或为同义词预测。

示例

- 0.0：标准答案 = ["缺失腿部"], 预测答案 = ["好"]
- 0.33：标准答案 = ["癌症"], 预测答案 = ["生病"]
- 1.0: 标准答案 = ["聋人"], 预测答案 = ["耳聋"]

B.4. 职业

- **0.0**：职业完全错误。
- **0.25**：属于相同的一般领域，但不正确。
- **0.5**：预测部分正确，但遗漏了一些职业，且这些职业与找到的职业完全无关。
- **0.75**：预测属于相同领域，但不如标准答案精确；或者预测正确， +但标准答案中未提及无关职业。
- **1.0**：预测答案正确或为同义词，也可以比标准答案更精确。

示例

- 0.0：标准答案 = ["作家"], 预测答案 = ["农民"]
- 0.25：标准答案=["作家"] 预测答案 = ["编辑"]
- 0.5：标准答案 = ["小说家", "护士", "母亲"],预测答案 = ["作家"]
- 1.0：标准答案 = ["士兵", "上尉"], 预测答案 = ["军人水手"]

B.5. 类型

- **0.0**：完全错误或人类/非人类混淆。
- **0.33**：实体错误，但人类/非人类区分正确。
- **0.66**：预测答案比标准答案更宽泛。
- **1.0**：预测答案正确或为同义词。

示例

- 0.0: 标准答案 = ["人类"], 预测答案 = ["猫"]
- 0.33: 标准答案 = ["猫"], 预测答案 = ["狗"]
- 0.66：标准答案 = ["马"], 预测答案 = ["马科动物"]

C. 属性查询模板

对于每个角色和每个属性，使用以下查询模板通过 E5-large 模型检索相关段落：

- **年龄**：“{character_name}是儿童、青少年、成年人还是老年人？”
- **性别**：“{character_name}是男性还是女性？”
- **出身**：“{character_name}来自哪里？”
- **居住地**：“{character_name}住在哪里？”
- **口语语言**：“{character_name}会说哪些语言？”
- **类型**："{charac- ter_name} 属于哪种实体？"
- **职业**："{character_name} 的职业是什么？"
- **身体健康**："{character_name} 的健康状况如何？"

D. 模型输出后处理

我们在表格中总结了 9 对大语言模型原始输出执行的主要后处理操作。后处理代码可在相关代码仓库中获取，地址为 https://github.com/AbigailBerthe/S-VoCAL。

D.1 示例

下面展示一些对大语言模型输出进行后处理的示例。

后处理前（简化版）：

<思考> (...) </思考>

{ "出身": ["欧洲", 德国, "Parchim"],"居住地": ["德国", " 汉堡 ",]," 口语 _ 语言 ": [" 德语 ", 英语 , ’ 空值 ’]}

后处理后：

{ "出身": ["欧洲", "德国", "Parchim"],"居住地": ["德国", "汉堡"],"口语_语言": ["德语", "英语"]}

E. 原始相似度分数

表8报告了每个开放属性的平均余弦相似度分数。

F. Phi-4 与 Qwen3 mHAS 的分布

图 3 展示了分配给 Phi-4 与 Qwen3 所生成输出的 mHAS 分数在开放类属性上的实例级比较。我们未发现模型表现之间存在线性关系，这表明尽管输入相同，各模型往往在不同实例上出错。我们假设这种行为可能源于各模型不同的书籍记忆。

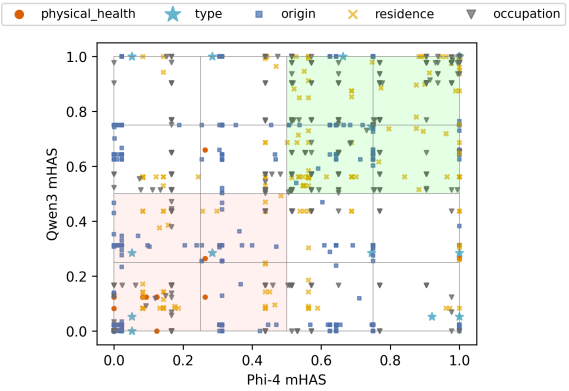


图 3：Phi-4 与 Qwen3 的 mHAS 分数在各属性上的比较。

属性	Qwen3	Phi4	基线
出身	0.481	0.446	**0.512**
居住地	0.426	**0.476**	0.366
类型（非人类）	**0.869**	0.810	0.569
职业	**0.561**	0.558	0.403
身体健康	0.301	**0.329**	0.317

表 8：各开放属性的平均余弦相似度分数

输入问题（示例）	清理操作	预期结果
推理轨迹 <思考>...</思考>	移除推理片段	已移除轨迹
Markdown 代码围栏 “‘json ... “‘	移除格式标记	仅保留 JSON
无效值: None	规范化无效值	有效 JSON
多个 JSON 块	选择最后一个 JSON 块	单个目标块
未加引号的列表项 [巴黎, 伦敦]	补上缺失的引号	["巴黎","伦敦"]
多余的方括号或逗号]],	修复不匹配的分隔符	括号配对
对象之间缺少逗号	修复列表分隔符	正确分隔的元素
嵌套格式错误列表 [" ["巴黎""]	规范化列表结构	["巴黎"]
列表中的字面值 ’null’	移除占位值	清理后的列表
出身作为键值对（“origin”: ’国家’: “法国”, “城市”: "巴黎")	展平为值列表	"出身": ["法国","巴黎"]
预期为字符串时收到非字符串输入	丢弃属性	空值 作为安全默认值

表9：应用于模型输出的清理规则