使用大型语言模型进行短语断句预测的合成数据生成

Hoyeon Lee[1], *Sejung Son*[2], *Ye-Eun Kang*[3], *Jong-Hwan Kim*[1]

[1]NAVER Cloud, South Korea
[2]NHN, South Korea
[3]Yale University, USA

yeon.lee@navercorp.com

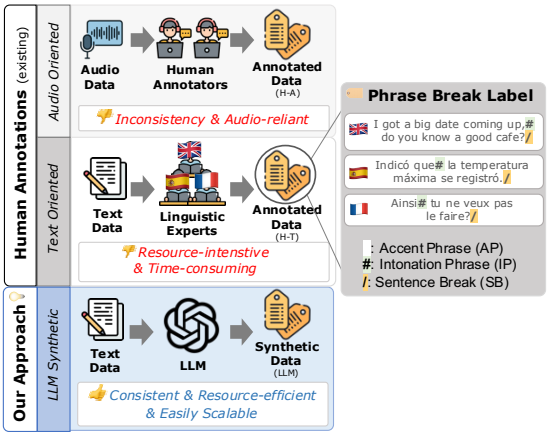

Figure 1: 传统人工标注和我们基于 *LLM* 生成的短语断句预测合成数据概述。

Abstract

当前的短语断句预测方法解决了语音合成系统中的关键韵律问题，但严重依赖于音频或文本的大量人工标注，从而招致显著的人力投入和成本。语音领域中由语音因素驱动的固有变异性，使得获得一致的高质量数据更加复杂。最近，大型语言模型（LLMs）在通过生成定制的合成数据来解决 NLP 中的数据挑战方面表现出成功，同时减少了对人工标注的需求。受此启发，我们探讨利用 LLM 生成合成短语断句标注，通过与传统标注进行比较并评估多语言效果，解决人工标注和语音相关任务的挑战。我们的研究结果表明，基于 LLM 的合成数据生成有效缓解了短语断句预测中的数据挑战，并突显了 LLM 作为解决语音领域问题的可行方案的潜力。

Index Terms: synthetic data generation, large language model, phrase break prediction, text-to-speech front-end, multilingual

1. 引言

文本到语音（TTS）本质上是一个一对多映射任务，因为一个文本输入可以根据说话人和韵律的变化产生多个语音输出。这种歧义通常导致在生成自然且情境适宜的语音时面临挑战。影响这种变化性的一个关键前端模块是短语断句预测，它决定了短语之间合适的停顿位置。准确的短语断句预测对于捕捉有助于语音自然性的细微韵律至关重要，特别是在句法和韵律规则在各语言之间显著不同的多语言 TTS 中 [1–3] 。为此，先前的研究利用人工标注作为定义短语边界和捕捉特定语言韵律变化的重要资源。

我们考虑了用于标注短语断点的两种不同的人类标注方法，如图 1 所示。在第一种方法中，人类标注者通过检测停顿直接从音频中识别短语断点。这种方法被称为面向音频的标注 [2, 3] 。在第二种方法中，标注者仅依靠文本，通过分析句子结构以确定短语断点的位置。这些标注被称为面向文本的标注 [4–6] 。尽管这些方法在以前的研究中被广泛使用，但每种方法都有其自身的一系列挑战。面向音频的标注依赖于大量高质量的音频数据，并容易受到录音条件变化和语音因素引起的差异影响 [7, 8] 。相比之下，面向文本的标注需要广泛的语言学专业知识，而且相当耗时。此外，由于这两种方法都依赖于人为的手工努力，构建大规模、高质量的短语断点数据集需要巨大的时间和财务资源。对人力投入的高度依赖推高了成本和复杂性，当开发多语言的短语断点预测时，这一负担更是加剧。同时，随着 LLMs 在 [9–11] 方面的显著进展，最近的研究集中在探索其潜力，并在一系列下游任务中展示了惊人的结果。在各种应用中，基于 LLM 的合成数据生成作为一种有前景的方法出现，以应对数据稀缺和相关挑战，证明其在生成针对目标任务的定制高质量数据集方面的有效性。然而，这些研究主要集中在文本任务上，如分类 [12] 和机器翻译 [13] ，而与语音相关的领域相对未被充分探索，尽管在数据需求和限制方面通常更具挑战性。在本文中，我们通过实验证实了基于 LLM 的合成数据生成在短语断点预测中的有效性，重点解决语音领域中手工标注的固有挑战。为此，我们利用 LLM 这一面向文本的模型，只使用极少量的样例，远少于传统方法中通常需要的数千例。另外，我们提出了基于跨语言知识转移的新策略，以研究这种方法是否能在保持最低资源需求的情况下扩展到多种语言。为了进一步评估其实用性，我们分别在人工和合成标注上训练了预训练的多语言模型 MiniLM [14] ，并分析其对短语断点任务的影响。我们的贡献总结如下：

- 我们进行了第一次关于 LLM 生成的短语断句标注的实证研究，将其与两种传统的人类标注在一致性和与人类判断的对齐方面进行比较。
- 我们的结果表明，LLM 生成的注释在质量上可与人工注释媲美，同时实现了稳定和成本高效的标记，所使用的样本显著减少，且自动化程度比现有方法更高。
- 我们的方法展示了很强的泛化能力，在多种语言中实现了稳健的结果，并验证了其在不同语言环境中的适用性。

2. 相关工作

2.1. 短语断句预测

大多数基于深度学习的语句断句模型 [15, 16] 通过利用大规模的人类标注表现出优于传统方法的性能 [17] 。这些标注通常是通过音频导向 [2, 3] 或文本导向 [4–6] 方法收集的：前者依赖于韵律提示，而后者使用文本句法。Futamata 等人 [2] 收集了 99,907 条针对单一语言的音频导向标注，这表明即使在单语环境中也需要大量的努力。尽管 [3] 探索了跨语言知识迁移以降低标注成本，其方法仍然需要为源语言提供超过 60,000 个人工标注。

Prompt

*You are an **English** linguistic expert, skilled in identifying phrase breaks when reading text out loud. **Your task is to label phrase break boundaries for the given input sentence given the following information.***

A phrase break in the input sentence is represented as follows:
*- **#**: This symbol is added after a word when a phonetic pause is needed while reading the sentence aloud.*
*- **/**: This symbol is added after a word at the end of a sentence, such as when it is followed by a period (.), an exclamation mark (!), or a question mark (?).*

***The desired output format for the input sentence** is the original sentence, with each word concatenated with #, /, or no symbol, while maintaining the original spacing between the words.*
The input sentence must not be altered in any way, and no word should have more than one break symbol attached.

For the following input, please return the desired output in the correct format, preceded by Desired Output:
Please provide only the direct answer, with no introductory phrases or analysis.
Remember to try speaking it aloud, before you annotate!

Here are some examples of inputs and their desired outputs.
Input: I got a big date coming up, do you know a good restaurant?
Desired Output: I got a big date coming up,# do you know a good restaurant?/
....

Based on the examples above, annotated this input.
Input: How will this new technology affect the world of commerce and mobile?

LLM

Desired Output: How will this new technology# affect the world of commerce and mobile?/

Figure 2: 用于生成短语断句注释的提示。

2.2. 使用 LLMs 生成合成数据

大型语言模型在各种自然语言处理任务中表现出色 [9–13]，尤其是在合成数据生成方面 [18,19]。它们有效地解决了数据相关的挑战，例如为文本分类增加标记数据 [12] 和生成机器翻译数据 [13]。然而，大多数基于 LLM 的合成数据生成的成功都局限于自然语言处理任务，使得语音领域相对未被充分探索。与文本不同，语音数据需要像节奏和音高这样的韵律特征，使得标注更加复杂。这些挑战在多语言设置中变得更加严重，因为不同的韵律结构和句法差异进一步复杂化了一致性标注。

3. 方法论

在本节中，我们概述了使用 LLM（即，先进的 GPT-4o mini [11]）生成短语断句标注时所遵循的程序。我们假设仅能访问有限的一组预定义短语断句标签。我们使用精心设计的系统提示（图 2），赋予 LLM 语言学专家的人格，同时通过任务提示指示它在不改变原始文本的情况下用“#”标注语音停顿，用“/”标注句子边界。为了提高准确性，LLM 的人格被设置为在标注前仔细阅读句子并大声朗读，模拟人类标注过程。

在第一项研究（第 4 节）中，我们对不同类型的英文注解进行了探索性分析，作为验证基于 LLM 方法用于短语断句生成潜力的初步步骤。具体来说，我们通过将 LLM 的注解（LLM）与两种不同的人类注解类型进行比较：以音频为导向的（H-A）和以文本为导向的（H-T），来评估 LLM 的执行指令并验证其注解的能力。基于这一分析，我们在第 5 节评估合成注解的质量。为了在英语之外验证基于 LLM 的合成注解，在英语资源通常丰富的情况下，我们将评估扩展到法语和西班牙语，这些语言的资源比英语少。随后，我们通过将 LLM 的角色从英语专家转换为多语言专家，并在少样本示例中加入英文示例（LLM 训练中最常用的语言），来调整提示以实现跨语言转移。最后，在第 6 節，我們通过在 LLM 生成的注解上训练一个更小的模型 [14]，來评估合成数据的实际价值，之后评估其性能以确定 LLM 注解是否能够切实替代人工注解。

4. 探索性分析：不同标注类型的比较

4.1. 实验装置

数据集我们从新闻、社区网站和口语语料库中收集了来自八位母语为英语话者的 1000 条英语话语，以捕捉不同的语音模式。这些话语被用两种方式进行人工标注：使用录音（H-A）和通过依赖文本的语言专家（H-T）。对于 H-A，标注者根据静音和转换标记休止符，遵循 [2,3]。每个话语都用三个短语休止符标签进行标注：重音短语（AP）、语调短语（IP）和句子边界（SB）。

Table 1: 按例子数量的标注比较（k）。通过率（%）和每个话语的平均措辞信息（%），括号内为标准差。

Annotation Type		Pass Rate	AP	IP	SB
Human	H-A	100.	79.0 (11.3)	9.2 (10.4)	11.9 (6.1)
	H-T	100.	79.3 (8.2)	8.8 (7.4)	12.0 (6.4)
LLM	ZS	98.1	18.4 (7.4)	72.2 (11.1)	9.4 (3.9)
	FS, k=2	99.3	86.1 (6.6)	4.6 (6.5)	9.3 (3.8)
	FS, k=4	98.5	84.8 (6.4)	5.9 (6.7)	9.3 (3.7)
	FS, k=8	98.4	83.1 (5.9)	7.6 (6.4)	9.3 (3.7)
	FS, k=16	98.4	82.3 (5.8)	8.5 (6.3)	9.3 (3.7)
	FS, k=32	98.5	83.2 (5.9)	7.5 (6.4)	9.2 (3.7)
	FS, k=64	98.4	82.9 (5.9)	7.9 (6.3)	9.2 (3.7)
	FS, k=128	98.3	83.2 (5.8)	7.5 (6.3)	9.3 (3.8)
	FS, k=256	98.1	82.3 (6.4)	8.0 (6.6)	9.7 (4.1)

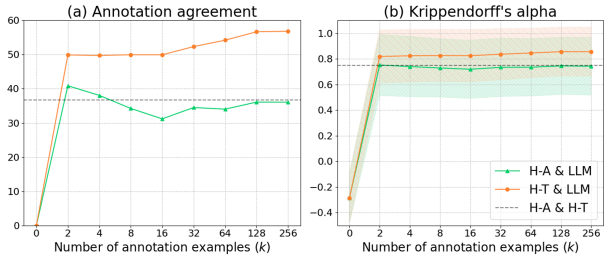

Figure 3: 按 k 分类的标注类型的协议（%）和 *Krippendorff's* α。阴影区域表示标准差。

模型 & 提示，我们使用 GPT-4o mini [1] 进行数据生成，基于其强大的整体性能和成本效益 [20,21]。通过专注于单一种表现优异的 LLM，我们旨在隔离不同标注类型的影响，从而能够更有控制地分析它们对短语停顿的影响。我们将温度设置为 0.0，top-p 设置为 1.0，并应用图 2 中的提示。然后，模型在零次（ZS）和少次（FS）设置中生成合成标注，使用来自同一领域的专家标注的语句以保持一致性。

指标我们使用通过率 [22] 来衡量模型在短语断句生成任务中遵循指令的程度。此外，我们使用一致性 [23] 和 Krippendorff 的 α [24] 来衡量相同话语的注释一致性。一致性验证序列中所有短语断句是否匹配，而 Krippendorff 的 α 则量化标注者之间的可靠性。

4.2. 结果与分析

4.2.1. 使用 LLM 生成合成数据

所提出的提示有效地引导 LLM 遵循指令。为了验证 LLM 是否正确地遵循了提示指令，我们首先检查了表 1 中的通过率。虽然短语断点标签的分布在 H-A 和 H-T 之间表现出相似的平均值，但 H-T 在 AP 和 IP 中的标准差较低，表明一致性更好。这种在 H-A 中的变化可以归因于录音条件和说话者特性上的差异 [25]。对于合成标注，所有设置下的通过率与人工标注相当，验证了我们提示的有效性。

少样本示例可以改善大型语言模型对人类标注的对齐。虽然通过率保持稳定，但措辞信息的分布显示出显著差异。在零样本设置中，合成标注与 H-A 和 H-T 明显偏离，这突显了在没有标注示例的情况下生成正确标签的困难。随着 k 的增加，这些值逐渐与人类标注对齐，最终与 H-A 和 H-T 趋同。这表明引入相关示例有助于大型语言模型更好地遵循特定任务的标注方法，从而提高一致性。

[1] gpt-4o-mini-2024-07-18

Table 2: 英语中人工注释和 *LLM* 注释的表现。使用 *H-A & H-T* 作为目标标签的人类评分 *& F1*，括号内为与人工注释的差距。

Annotation Type		Human Score	F1 (H-A)	F1 (H-T)
Human	H-A	56.30	-	82.14
	H-T	85.70	82.14	-
LLM	ZS	0	39.31(-42.8)	39.56(-42.6)
	FS, k=2	71.61	79.52(-2.6)	86.12(+4.0)
	FS, k=4	72.50	80.13(-2.0)	87.33(+5.2)
	FS, k=8	75.13	80.79(-1.4)	88.31(+6.2)
	FS, k=16	76.61	80.73(-1.4)	88.56(+6.4)
	FS, k=32	78.61	81.17(-1.0)	88.74(+6.6)
	FS, k=64	79.46	81.33(-0.8)	89.55(+7.4)
	FS, k=128	81.60	81.89(-0.3)	90.19(+8.1)
	FS, k=256	83.72	81.81(-0.3)	90.25(+8.1)

4.2.2. 不同标注类型的可靠性比较

在 H-A 和 H-T 中的不同的韵律线索导致了低一致性。我们进行了比较分析来检查标注类型之间的相关性，如图 3 所示。结果表明，尽管 H-A 和 H-T 常被用作基准真相，它们的一致性仍然相对较低，大约在 36–38 % 之间。这表明，虽然这两种方法都能捕捉到有意义的短语中断模式，但它们对韵律的强调不同——H-A 受到自然语音变化的影响 [7, 8] ，而 H-T 更符合句法结构 [5, 6] 。此外，Krippendorff 的 α 表明中等可靠性，这进一步强化了 H-A 和 H-T 并非完全可以互换的观点。它们的差异表明各自提供了互补的视角。

大语言模型（LLMs）能为语音任务生成类似人类的标签吗? 在零样本设置中，LLM 的标注显示与人类标注完全不一致（AG = 0, $\alpha \approx -0.28$ ）。我们的错误分析证实，零样本提示是无效的，因为 LLM 错误放置了断句符号，过度使用 " # "，或在标点符号周围错误断句。相比之下，即使提供少量示例，也能显著改善所有标注类型的一致性。虽然 H-A & LLM 在大多数 k 值下仍低于人类一致性，但 H-T & LLM 稳步改善，增加了 15-17 % 并达到了 $\alpha \geq 0.85$ ，如图 3 所示，表明了一种高度可靠的一致性水平。提供足够的示例后，LLM 标注与 H-T 紧密对齐，表明其能够从文本线索中捕捉结构性韵律模式，拓展了它们在语音任务中的潜力。

5. 评估 I：注释质量

5.1. 实验设置

数据集按照在第 4 节中用于英语数据的相同程序，我们分别为法语和西班牙语收集了 1,000 个语句，并附有两个对应的人类注释。相同的短语断点标签和领域适用于所有设置。

模型 & 提示大多数模型配置来自第 4 节，并进行了几个关键修改。考虑到在实际应用中获取特定语言少数示例的困难，我们保持 $k = 16$ 固定。我们排除了与人工标注显著偏差的零样本设置，并包括了从源语言转移知识的跨语言设置。在跨语言设置中，我们使用来自英语 (En) 作为源语言和目标语言 X 的标注示例。我们将来自 En 和 X 的标注示例总数设为 16。

评估标准我们引入了一种人为评分方法，由语言学专家对每个文本-注释对进行二元评估（可接受/不可接受）。为了避免偏见，这些评估者与创建少样本示例和人为注释的人不同。根据 [3] ，我们采用宏平均的 F1 分数。

5.2. 结果与分析

5.2.1. 人工与合成标注

H-T 作为一个更一致的基准。为了从人的角度评估标注质量，我们在表格 2 中评估了人类评分。结果显示 H-A 和 H-T 之间大约有 30 % 的差距，这并不表示优劣，而是反映了 H-T 与人类评价的一致性，而 H-A 则捕捉了语音的多样性。如在第 4.2.2 节中所述，H-A 受说话者特质如语调和语速的影响，导致不一致性。因此，H-T 较高的分数反映了其结构的一致性，这表明 H-T 适合结构化任务，而 H-A 则捕捉自然语音模式。

LLMs 如何生成类似人类的标签? 与 H-A 相比，随着提供的样本数量的增加，F1 值持续上升，接近 82.14。此外，当与 H-T 进行评估时，LLM 在 $k = 2$ 时超过了 H-A 的 F1 值 82.14，表明其与以文本为导向的注释更为一致。然而，在某个 k 之后，人类评分和 F1 分数都趋于平稳，F1 分数的增加达到饱和，而在更高的 k 时，人类评分略有下降。这些结果表明，少量样本示例有助于 LLM 推断短语断句的韵律线索，从而扩展其在语音任务中的适用性。LLM 生成的注释与 H-T 一致，同时与 H-A 保持合理的一致性，在文本和语音韵律间形成平衡。因此，选择和优化少量样本示例对于最大化性能至关重要，强调了在捕捉由语音驱动的变化时提示设计的重要性。

5.2.2. 其他语言的注释

语言差异影响跨语言数据生成。结合英语数据提高了通过率和人评分，同时存在语言特定的效果。在法语中，人评分在 En12 + X4 达到顶峰，显示了有效的知识转移（表 3 ）。即便在 En16 + X0 人评分降低到 79.50，但仍然超过了 H-A 的人评分。对于西班牙语，平衡的比例至关重要，最高的人评分出现在 En8 + X8 。过多的英语数据使人评分在 En16 + X0 降至 70.90。这些差异来自于语言因素。添加英语数据提高了标注质量，因为法语和英语在结构上具有相似性 [26, 27] 。相反，西班牙语是一种音节定时语言，和压力定时的英语不同，导致不同的短语中断模式 [28] 。虽然英语数据对西班牙语有益，但过度依赖扰乱了韵律对齐，降低了人评分。这种模式与 LLM 模型多种语言的方式一致，捕捉到共享的结构同时保持语言特定的区别 [29] 。在跨语言场景中，结合目标语言数据与英语数据比仅依赖目标语言数据更能提升标注质量。

6. 评估 II：注释对模型性能的影响

6.1. 实验设置

数据集我们为每种语言使用相同的人类和 LLM 注释（见章节 4 和 5 ）。鉴于现实世界的限制，我们将 k 设置为 16，对于资源丰富的英语设置为 256。对于跨语言少样本学习（XL-FS）设置，我们使用两种配置：平衡 $En8 + X8$ 和仅英语 $En16 + X0$ 。我们从 CommonVoice [30] 中收集生成输入的话语，这是一个广泛使用的开放域语音数据集，覆盖广泛。

训练我们将标注数据按 85:7.5:7.5 的比例分为训练、验证和测试集。对于短语断句预测，我们使用 MiniLM [2] ，这是一种能够从每种语言的标注中学习的蒸馏多语言模型。选择在测试集上获得最高宏 F1 分数的模型进行评估。两个人工标注，H-A [2, 3] 和 H-T [4–6] ，作为与 LLM 生成的标注进行比较的基线。

表格 4 展示了测试集上的宏观 F1 分数和 CommonVoice 上的人工评分。通过大语言模型（LLM）注释训练的模型在英语中取得了最高的 F1，表明较低的主观性和更大的一致性可以提升性能。同样，法语模型也达到了最高的 F1，确认了少样本提示的有效性。值得注意的是，在 En16 + Fr0 设置中，LLM 注释在 F1 方面超过了 H-A 和 H-T。对于西班牙语，$k = 16$ 模型取得了与人工注释相当的 F1。类似地，使用合成数据训练的模型在 CommonVoice 上评估时优于人工注释模型。总体而言，LLM 注释在 F1 上稳

[2] https://huggingface.co/microsoft/Multilingual-MiniLM-L12-H384

Table 3: 在人类和 *LLM* 生成的注释在不同跨语言少样本设置下的表现。

Annotation Type		X = French				X = Spanish			
		Pass Rate	Human Score	F1 (H-A)	F1 (H-T)	Pass Rate	Human Score	F1 (H-A)	F1 (H-T)
Human	H-A	100.	67.50	-	82.91	100.	53.60	-	73.56
	H-T	100.	88.20	82.91	-	100.	86.70	73.56	-
LLM	FS (X 16 → X)	91.1	81.75	77.19(-5.7)	83.74(+0.8)	98.8	89.87	77.14(+3.6)	77.28(+3.7)
	En 4 + X 12 → X	94.2	78.73	79.09(-3.8)	85.13(+2.2)	99.3	91.14	74.03(+0.5)	77.13(+3.6)
	En 8 + X 8 → X	97.0	84.07	81.27(-1.6)	86.43(+3.5)	99.7	91.14	72.98(-0.6)	76.96(+3.4)
	En 12 + X 4 → X	99.7	84.89	83.63(+0.7)	86.89(+4.0)	99.8	82.28	71.28(-2.3)	74.17(+0.6)
	En 16 + X 0 → X	100.	79.50	84.77(+1.9)	86.02(+3.1)	100.	70.90	71.55(-2.0)	73.48(-0.1)

Table 4: 评估经过训练的注释的影响。

Language	Annotation Type	F1	CommonVoice
English	H-A	82.55	58.5
	H-T	89.89	72.2
	LLM (FS, k=16)	94.90	64.6
	LLM (FS, k=256)	94.51	75.7
French	H-A	90.32	76.2
	H-T	91.93	82.8
	LLM (FS, k=16)	98.59	69.0
	LLM (XL-FS, En 8 + Fr 8)	96.09	78.4
	LLM (XL-FS, En 16 + Fr 0)	94.06	84.5
Spanish	H-A	89.25	55.1
	H-T	89.57	56.1
	LLM (FS, k=16)	89.84	69.4
	LLM (XL-FS, En 8 + Es 8)	88.33	69.4
	LLM (XL-FS, En 16 + Es 0)	89.04	57.1

定地优于人工标注的数据，验证了少样本提示能够有效生成短语断点。此外，结果表明英语示例为短语断点生成提供了有意义的支持，加强了第 5.2.2 节中的发现。值得注意的是，这种方法证明是可推广的，在法语和西班牙语中取得了高性能，确认了其跨语言的适用性。

在本文中，我们对短语断句预测的合成数据生成进行了全面研究，利用大型语言模型（LLMs）解决语音领域由于语音因素引起的基本挑战。我们的结果表明，合成的短语断句标注有效地克服了传统人工标注的局限，这些局限由于需大量资源和人的参与导致的不一致而受到阻碍。我们还发现，使用合成数据训练的模型可以达到与那些使用从音频和文本来源得到的人类标注进行训练的模型相当甚至更优的表现。我们的工作突显了在资源有限的语言环境中，以最少量的例子进行低成本数据生成的可能性，提供了一种有前景的解决方案。

7. References

[1] Y. Xiao, L. He, H. Ming, and F. K. Soong, "Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6704–6708.

[2] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis," in *Interspeech 2021*. ISCA, 2021, p. 3126–3130.

[3] H. Lee, H.-W. Yoon, J.-H. Kim, and J.-M. Kim, "Cross-lingual transfer learning for phrase break prediction with multilingual language model," in *Interspeech 2023*. ISCA, 2023, p. 611–615.

[4] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech & Language*, vol. 12, no. 2, pp. 99–117, 1998.

[5] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading tts: Exploiting text structure information," in *Interspeech 2017*, 2017, pp. 1064–1068.

[6] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[7] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, "Nn-based ordinal regression for assessing fluency of esl speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7420–7424.

[8] H. Meng, W.-K. Lo, A. M. Harrison, P. Lee, K.-H. Wong, W.-K. Leung, and F. Meng, "Development of automatic speech recognition and synthesis technologies to support chinese learners of english: The cuhk experience," *Proc. APSIPA ASC*, pp. 811–820, 2010.

[9] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le *et al.*, "Least-to-most prompting enables complex reasoning in large language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 443–10 461.

[13] Q. Bolding, B. Liao, B. Denis, J. Luo, and C. Monz, "Ask language model to clean your noisy translation data," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, Dec. 2023, pp. 3215–3236.

[14] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020.

[15] S. Pascual and A. Bonafonte, "Prosodic break prediction with rnns," in *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*. Springer, 2016, pp. 64–72.

[16] A. Vadapalli and S. V. Gangashetty, "An investigation of recurrent neural network architectures using word embeddings for phrase break prediction." in *Interspeech*, 2016, pp. 2308–2312.

[17] H. Schmid and M. Atterer, "New statistical methods for phrase break prediction," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 659–665.

[18] S. Wang, Y. Liu, Y. Xu, C. Zhu, , and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4195–4205.

[19] B. Ding, C. Qin, L. Liu, Y. Chia, B. Li, S. Joty, and L. Bing, "Is GPT-3 a good data annotator?" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 11 173–11 195.

[20] T. Wang, Z. He, W.-Y. Yu, X. Fu, and X. Han, "Large language models are good multi-lingual learners: When llms meet cross-lingual prompts," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4442–4456.

[21] J. A. Schnabel, J. R. Trippas, F. Scholer, and D. Hettiachchi, "Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment," *arXiv preprint arXiv:2501.14296*, 2025.

[22] S. Iskander, S. Tolmach, O. Shapira, N. Cohen, and Z. Karnin, "Quality matters: Evaluating synthetic data for tool-using llms," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 4958–4976.

[23] J. Cheng, Y. Lu, X. Gu, P. Ke, X. Liu, Y. Dong, H. Wang, J. Tang, and M. Huang, "Autodetect: Towards a unified framework for automated weakness detection in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 6786–6803.

[24] A. Elangovan, J. Ko, L. Xu, M. Elyasi, L. Liu, S. Bodapati, and D. Roth, "Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge," *arXiv preprint arXiv:2410.03775*, 2024.

[25] A. M. Kapadia, J. A. Tin, and T. K. Perrachione, "Multiple sources of acoustic variation affect speech processing efficiency," *The Journal of the Acoustical Society of America*, vol. 153, no. 1, pp. 209–209, 2023.

[26] Y. Liu, "A comparative study of english tense and french tense." *Theory & Practice in Language Studies*, vol. 4, no. 11, 2014.

[27] J. Vander Klok, H. Goad, and M. Wagner, "Prosodic focus in english vs. french: A scope account," *Glossa: a journal of general linguistics*, vol. 3, no. 1, 2018.

[28] L. Colantoni, O. Marasco, J. Steele, and S. Sunara, "Learning to realize prosodic prominence in l2 french and spanish," in *Selected proceedings of the 2012 second language research forum: Building bridges between disciplines*, 2014, pp. 15–29.

[29] R. Zhang, Q. Yu, M. Zang, C. Eickhoff, and E. Pavlick, "The same but different: Structural similarities and differences in multilingual language modeling," *arXiv preprint arXiv:2410.09223*, 2024.

[30] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, May 2020.