# TICL：文本嵌入 KNN 用于语音上下文学习，解锁大型多模态模型的语音识别能力

Haolong Zheng, Yekaterina Yegorova, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign
{haolong2, yay2, jhasegaw}@illinois.edu

ABSTRACT

语音基础模型最近展示了执行语音上下文学习(SICL)的能力。选择有效的上下文示例对于 SICL 性能至关重要，然而，选择方法仍然缺乏探索。在这项工作中，我们提出了用于 SICL 的文本嵌入 KNN（TICL），这是一个简单的流程，利用语义上下文来增强现成的大规模多模态模型的语音识别能力而无需微调。在包括口音英语、多语言语音和儿童语音在内的具有挑战性的自动语音识别任务中，我们的方法使模型超越零样本性能，最多相对 WER 减少了 84.7%。进行了消融研究以展示我们方法的鲁棒性和效率。

Index Terms— 上下文学习，自动语音识别，大型多模态模型

## 1. 介绍

上下文学习 (ICL) [1] 已被广泛应用于大型语言模型 (LLMs)，通过在输入上下文中对提供的演示进行条件设置，从而适应新任务 [2]。它避免了代价高昂的微调，后者可能导致灾难性遗忘问题。影响 ICL 性能的关键因素是上下文示例的选择 [3–5]。

语音情境学习（SICL）将这一想法扩展到可以处理语音输入的模型，这最初是在 [6] 中提出的。它应用 SICL 来增强 Whisper 对中文方言的 [7] 识别能力，在此过程中使用 Whisper 嵌入作为情境示例检索 k 最近邻。它们通过搜索最相似的说话人嵌入来减少每个样本的候选数据集。他们的实验表明，选择情境示例会影响 SICL 的性能。然而，由于计算复杂性，实验仅限于词级语音。为了解决这一问题，[8] 提出了一个句子级别的数据存储库，在其中，话语嵌入被平均池化并在时间无关的空间中检索。尽管如此，Whisper 相对较小的上下文窗口限制了能够有效整合的示例数量。

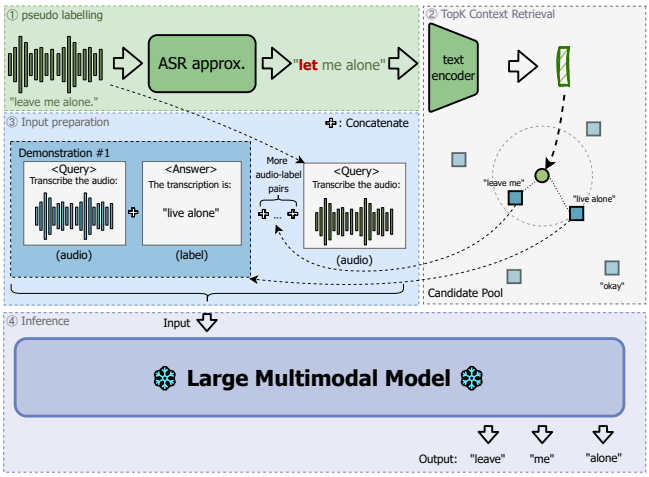


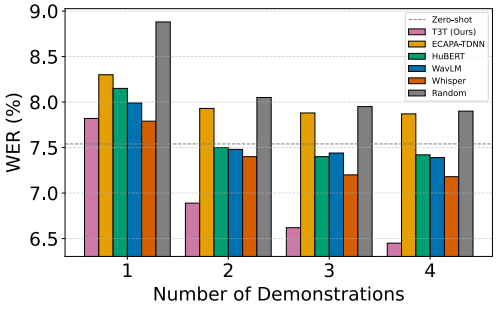

Fig. 1. TICL **管道概述**: (1) 预训练的 ASR 为测试音频生成伪标签。(2) 文本编码器嵌入这个伪标签，并从一组预嵌入了相同编码器真实转录的池中检索出最接近的 $K$ 个候选者。(3) 检索到的 (音频, 标签) 对被作为上下文学习示例添加在测试音频之前。(4) 示例和测试音频被传递为输入给大型多模态模型（LMM）以生成最终转录输出。

随着大型多模态模型（LMMs）提供更大上下文窗口的出现，SICL 研究得到了扩展。例如，[9] 使用了 Phi-4-Multimodal-instruct [10] 并将演示的数量扩大到 12 个。这在带口音的英语语音识别方面带来了相对 19.7%的性能提升。[11] 评估了 Gemini-1.5 Pro [12] 在带有多达 100 个演示的带口音的语音上的 SICL 能力。然而，这两种方法都依赖于随机抽样来获取 SICL

Table 1. TICL 带声调英语结果。↓错误率%

| | GLOBE-V2 | | L2-**北极区** | |
|---|---|---|---|---|
| | Phi-4-MM | Qwen2-**音频** | Phi-4-MM | Qwen2-**音频** |
| $\boldsymbol{k}$**=0** | 4.23 | 5.41 | 8.47 | 11.06 |
| $k$=1 | 1.40 | 2.38 | 2.94 | 4.76 |
| $k$=2 | 1.13 | 2.67 | 2.81 | 1.7 |
| $k$=3 | 0.92 | 1.89 | 2.70 | 1.52 |
| $k$=4 | 0.88 | 1.66 | 2.62 | 1.41 |
| $\Delta_{\text{rel}}$ | 79.2% | 69.3% | 69.1% | 84.7% |

Fig. 2. 检索方法比较结果




示例，这可能会低估 SICL 的潜力。

为了解决选择 SICL 示例的挑战，我们引入了用于 SICL 的文本嵌入 KNN（TICL），这是一个旨在利用高质量演示文稿语义检索的框架，如图 1 所示。

## 2. 方法论

### 2.1. 语音上下文学习

ICL 可以通过在包含目标领域数据的演示上进行条件设置来调整模型，而不是更新模型权重。与纯粹基于文本的 ICL 不同，SICL 同时对音频和文本标记进行条件设置。对于适应的 ASR，给定一个测试语音样本 $\boldsymbol{s}^*$，模型 $\Lambda$ 生成预测转录 $\hat{\boldsymbol{y}}$ 如下：

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \Pr(\boldsymbol{y} \mid C, \boldsymbol{x}_{\boldsymbol{s}}^*, \Lambda),$$

其中 $\boldsymbol{x}_{\boldsymbol{s}}^*$ 对应于 $\boldsymbol{s}^*$ 的音频编码，而 $C$ 表示上下文通常，$C$ 结构为查询-答案对 $c^{(i)} = (q^{(i)}, a^{(i)})$ 的对话历史 [1]。在我们的案例中，查询 $q$ 是由一个编码的文本提示 $\boldsymbol{x}_{\boldsymbol{p}}^{(\boldsymbol{i})}$ 和一个编码的音频片段 $\boldsymbol{x}_{\boldsymbol{s}}^{(\boldsymbol{i})}$ 连接而成，而答案 $a$ 是对 $\boldsymbol{s}^{(\boldsymbol{i})}$ 的转录，表示为 $\boldsymbol{y}^{(\boldsymbol{i})}$。因此，稍微滥用一下符号，我们案例中的完整上下文 $C$ 可以表示为：

$$C = [c^{(1)}; c^{(2)}; \ldots; c^{(n)}], \quad c^{(i)} = [\boldsymbol{x}_{\boldsymbol{p}}^{(\boldsymbol{i})}; \boldsymbol{x}_{\boldsymbol{s}}^{(\boldsymbol{i})}; \boldsymbol{y}^{(\boldsymbol{i})}].$$

### 2.2. 文本嵌入 K 近邻候选选择

本工作的目标是为给定的测试样本找到一个适当的上下文 $C$。[13,14] 表明，当 ICL 示例与测试样本属于同一领域时，模型从该上下文中获益最多。受到这一动机的启发，我们通过选择与测试音频的文字记录 $\boldsymbol{y}^*$ 在词汇上相似的转录文字 $\boldsymbol{y}^{(\boldsymbol{i})}$ 的样本来选取 SICL 示例。为了在词汇空间中获得最接近的样本，我们使用预训练的文本编码器对转录进行编码，并利用欧几里得距离检索出 $K$ 个最近邻候选者。然而，在推理过程中，$\boldsymbol{y}^*$ 不可用。因此，我们使用一个预训练的 ASR 模型生成伪标签 $\tilde{\boldsymbol{y}}$，该伪标签可能包含错误，然后我们将伪标签编码以获得近似的嵌入。

更具体地说：令候选数据集为 $\mathcal{C} = \{(\boldsymbol{s}^{(\boldsymbol{i})}, \boldsymbol{y}^{(\boldsymbol{i})})\}_{i=1}^{N}$，其中 $\boldsymbol{s}^{(\boldsymbol{i})}$ 和 $\boldsymbol{y}^{(\boldsymbol{i})}$ 分别是语音音频及其转录。令 $\phi : \mathcal{Y}_{\text{text}} \to \mathbb{R}^d$ 为一个冻结的句子嵌入模型，该模型将所有句子集合 $\mathcal{Y}_{\text{text}}$ 中的任何文本句子映射到一个 $d-$维向量。我们将 $\bar{\phi}(\boldsymbol{y}) = \phi(\boldsymbol{y})/\|\phi(\boldsymbol{y})\|_2$ 定义为 $\ell_2$ 归一化的嵌入。对于每个候选者 $c \in \mathcal{C}$，我们预计算 $\bar{\boldsymbol{z}} = \bar{\phi}(\boldsymbol{y})$。令 $f_\theta : \mathcal{X}_{\text{audio}} \to \mathcal{Y}_{\text{text}}$ 表示一个冻结的 ASR 模型（例如，Whisper）。在推理过程中，给定一个音频 $\boldsymbol{s}^*$，我们首先获得伪转录 $\tilde{\boldsymbol{y}} = f_\theta(\boldsymbol{s}^*)$ 并计算其词汇嵌入 $\bar{\boldsymbol{z}}^* = \bar{\phi}(\tilde{\boldsymbol{y}})$。我们选择在归一化嵌入空间中使用欧几里得距离选出的 $K$ 个最近候选者，

$$\mathcal{N}_K(\boldsymbol{s}^*) = \text{TopK}_{i \in [N]}\big(-r(i)\big),$$

其中 $r(i) = \big\|\bar{\boldsymbol{z}}^* - \bar{\boldsymbol{z}}^{(\boldsymbol{i})}\big\|_2$. 最后，使用这 $K$ 个候选者根据 2.1 构建上下文 $C$。

## 3. 实验

所提出的流水线设计适用于各种语音识别任务，并且与任何 LMM 兼容。在这项工作中，我们将它应用于 Phi-4-MultiModal-instruct (Phi-4-MM) [10] 和部

[1]初步实验表明，如果我们建模 $\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P_{\text{SICL}}(\boldsymbol{y} \mid \boldsymbol{X}_{\boldsymbol{a}}, \boldsymbol{Y}_{\boldsymbol{c}}, \Lambda)$ 其中的 $\boldsymbol{X}_{\boldsymbol{a}} = [\boldsymbol{x}_{\boldsymbol{a}}^{(1)}; \ldots; \boldsymbol{x}_{\boldsymbol{a}}^{(\boldsymbol{i})}; \boldsymbol{x}_{\boldsymbol{a}}^*]$ 和 $\boldsymbol{Y}_{\boldsymbol{c}} = [\boldsymbol{y}^{(1)}; \ldots; \boldsymbol{y}^{(\boldsymbol{i})}]$，将会发生严重的幻觉。

Table 2. TICL 多语言结果对于 Phi-4-MM。默认情况下为↓WER%，而对于 zh/ja/th 则为↓CER%。

| | 本机支持的语言 | | | | | | | | 不受支持的语言 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 的 | 输入 | 是的 | 法语 | 它 | Ja | 点 | zh | 不列表 | 计划 | 鲁 | 个 | 痕迹 |
| ***k*=0** | 5.24 | 7.56 | 4.27 | 8.00 | 3.79 | 13.00 | 6.06 | 8.49 | 101.15 | 117.55 | 122.75 | 134.21 | 132.74 |
| $k$=1 | 8.82 | 7.82 | 6.66 | 8.08 | 4.79 | 8.11 | 3.60 | 13.15 | 58.41 | 37.91 | 21.49 | 68.23 | 39.87 |
| $k$=2 | 6.16 | 6.89 | 5.78 | 7.63 | 4.09 | 7.06 | 3.73 | 12.67 | 58.37 | 35.58 | 19.82 | 64.51 | 36.34 |
| $k$=3 | 5.75 | 6.62 | 5.74 | 7.48 | 3.75 | 6.38 | 3.63 | 11.86 | 60.78 | 35.66 | 18.86 | 65.42 | 36.47 |
| $k$=4 | 5.45 | 6.45 | 5.63 | 7.41 | 3.64 | 6.17 | 3.52 | 11.07 | 63.10 | 37.22 | 20.74 | 65.78 | 37.15 |
| $\Delta_{\text{rel}}$ | -4.0% | 14.7% | -31.9% | 7.4% | 4.0% | 52.5% | 41.9% | -30.4% | 42.3% | 69.7% | 84.6% | 51.9% | 72.6% |

分 Qwen2-Audio-7B-Instruct (Qwen2-Audio) [15]。除非另有说明，否则我们使用 Whisper-Large-v3-turbo 作为伪标签 ASR，因为它具有高准确性和快速运行时间。对于词汇检索，我们使用 all-mpnet-base-v2 为每个转录本表示一个句子嵌入。[2] 由于语义相似的句子在嵌入空间中彼此接近，因此即使存在小的伪标签缺陷，它也能恢复短语和意图匹配的示例。在多语言设置中，我们切换到重述 -多语言 -mpnet -基础 -v2[3] 以在一种与语言无关的空间中保持这些属性。默认情况下，除非另有指定，否则我们使用词错误率（WER）进行评估。我们的实验包括带口音的英语、多语言语音和儿童语音的数据集 [4]。所有结果都在官方测试分割上报告。

### 3.1. 检索方法比较

我们系统地比较了 Common Voice 英语子集上的不同检索策略。该数据集因其广泛的口音、话题、说话风格和录音条件而被选中，提供了大量且多样的检索池。除了两个基线方法外，即 Whisper-embedding 检索方法 [8] 和均匀随机选择上下文示例 [9]，我们还评估了捕获不同相似性维度的嵌入。具体来说，我们包括 HuBERT [16]（内容导向，近似语音/词汇相似性）、ECAPA-TDNN [17]（相对内容不变的说话人身份嵌入）和 WavLM [18]（结合身份和内容线索的目标说话人嵌入，并抑制环境噪声和干扰语音）。对于 HuBERT 和 WavLM，我们使用最后一层隐藏状态作为话语嵌入。对于具有时间维度的嵌入，我们在时间上应用统计池化以生成固定长度的向量。

图 2 显示 TICL 在所有设置中始终表现出最佳性能。基于内容和语音的检索方法（Whisper、HuBERT、WavLM）优于基于说话人的检索（ECAPA-TDNN），而随机基线表现最差。结果再次证实了上下文的质量可以影响 SICL 的性能。

### 3.2. 带声调的英语语音识别

对于带有口音的英语语音识别,我们使用 GLOBE-V2 [19], 它包含 164 种不同的口音, 以及 L2-ARCTIC [20]，它包含来自六种 L1 背景的非母语英语读者的语音。上下文示例是从训练和验证集中检索的。如表 1 所示，我们的 TICL 检索方法显示出普遍的和一致的改进，与零样本相比，最高可达 84.7%。

### 3.3. 多语言语音识别

对于多语言语音识别，我们使用了 Common Voice Corpus 15.0 数据集 [21]。为了评估对 Phi-4-MM 原生 ASR 能力的影响，我们重点关注直接支持的语言：英语（en）、中文（zh）、德语（de）、法语（fr）、意大利语（it）、日语（ja）、西班牙语（es）和葡萄牙语（pt）。此外，还包括俄语（ru）、荷兰语（nl）、波兰语（pl）、泰语（th）和土耳其语（tr），这些语言 Phi-4-MM 的 ASR 不原生支持。上下文示例是从验证分割中检索的。

从表 2 中，我们的 TICL 流水线对支持的语言是有效的。我们看到 ja 和 pt 有显著改善。然而，de、es 和 zh 的表现变差了。通过手动重新审视 zh 的结果，我们发现许多检索到的上下文示例在词汇上与目标文本不接近。两个因素导致这种情况：(i) 数据集包含复

[2] all-mpnet-base-v2

[3] 多语言 mppnet 基础版本 2.paraphrase

[4] 实验仅限于时长在 1 至 15 秒之间的语句。

Table 3. TICL 儿童语音结果为菲-4-MM。↓错误率%。

| | MyST | OGI | ENNI | RSR |
|---|---|---|---|---|
| **$k$=0** | 12.81 | 16.17 | 14.37 | 20.06 |
| $k$=1 | 17.27 | 9.55 | 17.57 | 18.92 |
| $k$=2 | 11.77 | 8.94 | 14.07 | 18.92 |
| $k$=3 | 11.69 | 8.75 | 13.54 | 18.90 |
| $k$=4 | 11.81 | 8.52 | 13.75 | 19.54 |
| $\Delta_{\text{rel}}$ | 8.7% | 47.3% | 5.8% | 5.8% |

合/罕见术语，伪标签生成器通常会产生带有重要错误的大致转录，从而降低词汇查询质量；(ii) 句子级多语言-MPNet 嵌入未能充分表示字符/子词的重叠，导致检索器错过接近匹配并转向无关示例。

表 2 进一步证明了我们的 TICL 管道使模型能够转录原本不支持的语言。在所有不支持的语言中都观察到了显著的改进，特别是对于 ru、tr 和 pl 有较强的提升。这些结果表明，在适当的设置下，SICL 可以解锁模型在未见过任务上的能力。

### 3.4. 儿童语音识别

为了评估儿童的语音识别性能，我们使用了四个语料库：My Science Tutor (MyST) [22]、OGI Kids' Speech Corpus [23]、Edmonton Narrative Norms Instrument (ENNI) [24] 和 Redmond Sentence Recall (RSR) [25]。我们按照 [26] 对 MyST 和 OGI 进行预处理，按照 [27] 对 ENNI 进行预处理。上下文示例是从训练集和验证集分割中检索出来的。如表 3 所示，TICL 在所有数据集上相对于零样本基线提供了稳定的改进。最大的改进是在 OGI 上，可能是由于候选池更好地匹配了需求。对于 MyST、ENNI 和 RSR，收益较为温和，我们将其归因于较小或多样性较低的候选池。

## 4. 消融研究

我们进行了两项消融研究，以分析伪标签器准确性以及上下文示例数量对我们提出的流水线的影响。我们使用了 GLOBE-V2 数据集，该数据集在规模和质量之间提供了良好的平衡。为了确保我们的发现适用于不同的 LMM 架构，我们同时使用 Phi-4-MM 和 Qwen2-Audio 进行了评估。

### 4.1. 伪标签器的敏感性

为了评估伪标签质量如何影响 TICL 管道，我们模拟了一系列的伪标签质量。伪标签使用不同大小（tiny、base、small、medium、large、large-v3-turbo）的 Whisper 配置生成，得到的伪标签 WER 从 13.11% 下降到 1.95%。对于每个设置，TICL 检索 $K = 4$ 个上下文示例，并在所有选定的语言模型上运行。如图 4.1 所示，高质量的伪标签始终表现出更好的性能。值得注意的是，即使是最嘈杂的标签，TICL 仍然大幅超越零样本基准线，而最佳和最差的伪标签生成配置之间的整体性能差距保持适度。这表明 TICL 对伪标签器的敏感度较低。我们将其稳健性归因于嵌入空间中的词汇检索：近义字符串（例如，“let me alone” 与 “leave me alone”）仍然接近，使检索能够减轻转录噪声。因此，尽管伪标签生成者引入了一个潜在瓶颈，但它对 TICL 的影响有限。

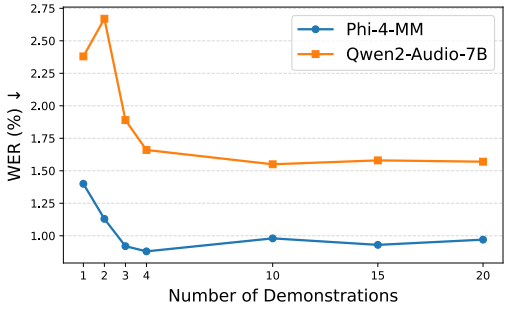
### 4.2. 上下文示例数量

为了研究更多的演示是否有所帮助，我们也评估了我们的管道与 $K \in \{1, 2, 3, 4, 10, 15, 20\}$ 的情况。如图 3 所示，当 $K$ 超过 $K \approx 4$ 时，增加的内容几乎没有带来益处，并且可能会影响准确性。我们相信有两个影响因素：(i) TICL 已经高效地找到了最有用的、语义上相近的例子，因此后来添加的内容更多是噪音而非信号；(ii) 音频的帧率高于文本输入，长提示会降低 LMMs 的性能，因为这些模型主要是在较短序列上训

Table 4. 伪标签器质量对 $K$=4 用于 GLOBE-V2 的 TICL 的影响。↓WER%。括号显示与相同模型的零样本学习 ($k$=0) 相比的相对 WER 降低幅度。

| **耳语会议** | **伪标签 错误率** | Phi-4MM | Qwen2-**音频**-7B |
|---|---|---|---|
| $k$=0 | – | 4.23 | 5.41 |
| tiny | 13.11 | 1.36 (↓ 67.85%) | 2.37 (↓ 56.19%) |
| base | 8.83 | 1.22 (↓ 71.16%) | 2.14 (↓ 60.44%) |
| small | 4.64 | 1.10 (↓ 74.00%) | 1.91 (↓ 64.70%) |
| medium | 2.86 | 1.01 (↓ 76.12%) | 1.88 (↓ 65.25%) |
| large | 2.17 | 0.92 (↓ 78.25%) | 1.82 (↓ 66.36%) |
| large-v3-turbo | 1.95 | 0.88 (↓ 79.20%) | 1.66 (↓ 69.32%) |

Fig. 3. 演示次数对 T3T SICL 管道的影响。在 GLOBE-V2 数据集上进行评估。




练的。实际上，TICL 管道只需少量演示即可很好地工作。

## 5. 结论

在本研究中，我们提出了一种增强且可泛化的 SICL 管道，称为 TICL。我们的方法在三项不同的语音识别任务中取得了显著的性能提升，最高达 84.7% 的相对改进。我们进行了消融研究，展示了我们提出的检索方法的鲁棒性和效率。本工作提供了一个轻量级且成本效益高的替代方案，用于将模型适应新领域，特别是对于 ASR。然而，当数据集包含罕见术语或复杂词汇时，我们观察到了一些局限性。未来的工作将探索解决这些失败情况的策略，并研究 SICL 范式的机制。

## 6. 致谢

本工作得到了国家科学基金会拨款#2229873 的支持。本工作使用了国家超级计算应用中心的 Delta 系统，该系统通过先进网络基础设施协调生态系统：服务与支持（ACCESS）计划下的 beiq-delta-gpu 分配获得。ACCESS 计划获得了国家科学基金会拨款#2138259、#2138286、#2138307、#2137603 和#2138296 的支持。

## 7. REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 1877–1901, Curran Associates, Inc.

[2] Qingxiu Dong, Lei Li, Damai Dai, et al., "A survey on in-context learning," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 1107–1128, Association for Computational Linguistics.

[3] Zihao Zhao, Eric Wallace, Shi Feng, et al., "Calibrate before use: Improving few-shot performance of language models," in Proceedings of the 38th International Conference on Machine Learning, Marina Meila and Tong Zhang, Eds. 18–24 Jul 2021, vol. 139 of Proceedings of Machine Learning Research, pp. 12697–12706, PMLR.

[4] Zhao Yang, Yuanzhe Zhang, Dianbo Sui, et al., "Representative demonstration selection for in-context learning with two-stage determinantal point process," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Houda Bouamor, Juan Pino, and Kalika Bali, Eds., Singapore, Dec. 2023, pp. 5443–5456, Association for Computational Linguistics.

[5] Rishabh Agarwal, Avi Singh, Lei Zhang, et al., "Many-shot in-context learning," in Advances in Neural Information Processing Systems, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds. 2024, vol. 37, pp. 76930–76966, Curran Associates, Inc.

[6] Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang, "Can whisper perform speech-based in-context learning?," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 13421–13425.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[8] Jiaming Zhou, Shiwan Zhao, Jiabei He, et al., "M2R-Whisper: Multi-stage and multi-scale retrieval augmentation for enhancing whisper," 2025.

[9] Nathan Roll, Calbert Graham, Yuka Tatsumi, et al., "In-context learning boosts speech recognition via human-like adaptation to speakers and language varieties," 2025.

[10] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al., "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025.

[11] Jian Cheng and Sam Nguyen, "Speech few-shot learning for language learners' speech recognition," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.

[12] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," arXiv preprint arXiv:2403.05530, 2024.

[13] Suzanna Sia and Kevin Duh, "In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models," in Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, Masao Utiyama and Rui Wang, Eds., Macau

SAR, China, Sept. 2023, pp. 173–185, Asia-Pacific Association for Machine Translation.

[14] Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden, "In-context example selection via similarity search improves low-resource machine translation," in Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 1222–1252.

[15] Yunfei Chu, Jin Xu, Qian Yang, et al., "Qwen2-audio technical report," arXiv preprint arXiv:2407.10759, 2024.

[16] Wei-Ning Hsu, Benjamin Bolte, et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[17] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Interspeech 2020. Oct. 2020, interspeech_2020, ISCA.

[18] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505 – 1518, Oct. 2022.

[19] Wenbin Wang, Yang Song, and Sanjay Jha, "GLOBE: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech," in Proc. Interspeech 2024, 2024, pp. 1365–1369.

[20] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, et al., "L2-ARCTIC: A non-native english speech corpus," in Interspeech 2018, 2018, pp. 2783–2787.

[21] R. Ardila, M. Branson, K. Davis, et al., "Common voice: A massively-multilingual speech corpus," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[22] Sameer Pradhan, Ronald Cole, and Wayne Ward, "My science tutor (MyST)–a large corpus of children' s conversational speech," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 12040–12045.

[23] Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole, "The OGI kids' speech corpus and recognizers," 10 2000, pp. 258–261.

[24] Phyllis Schneider, Denyse Hayward, Rita Vis Dubé, et al., "Storytelling from pictures using the edmonton narrative norms instrument," Journal of speech language pathology and audiology, vol. 30, no. 4, pp. 224, 2006.

[25] AI4ExceptionalEd, "Redmond sentence recall (RSR)," https://huggingface.co/datasets/ai4exceptionaled/Redmond-Sentence-Recall, Hugging Face dataset; accessed 2025-09-08.

[26] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan, "Benchmarking children's asr with supervised and self-supervised speech foundation models," 09 2024, pp. 5173–5177.

[27] Dancheng Liu and Jinjun Xiong, "FASA: a flexible and automatic speech aligner for extracting high-quality aligned children speech data," 2024.