倾听、想象与完善：一个基于大语言模型的启发式优化 ASR 校正框架

Yutong Liu[1], Ziyue Zhang[1], Yongbin Yu[1,⋆], Xiangxiang Wang[1,⋆], Yuqing Cai[1], Nyima Tashi[1,2]

[1] School of Information and Software Engineering,
University of Electronic Science and Technology of China, Chengdu, China.
[2] School of Information Science and Technology, Tibet University, Lhasa, China.
ybyu@uestc.edu.cn, xxwang@uestc.edu.cn

ABSTRACT

自动语音识别（ASR）系统仍然容易出现影响下游应用的错误。在本文中，我们提出了LIR-ASR，一个利用大语言模型（LLMs）并通过人类听觉感知启发而优化的迭代修正框架。LIR-ASR 采用了“倾听—想象—精炼”的策略，在上下文中生成音素变体并对其进行精炼。引入了带有有限状态机（FSM）的启发式优化，以防止校正过程陷入局部最优，并且基于规则的约束帮助保持语义保真度。在英语和中文 ASR 输出上的实验表明，与基线相比，LIR-ASR 实现了平均字符错误率/词错误率降低多达 1.5 个百分点，显示了转录准确性的显著提升。

Index Terms— 自动语音识别校正，大型语言模型，启发式优化

1. 介绍

大型音频模型 [1, 2, 3, 4] 由于其出色的鲁棒性、跨语言泛化能力和应对多种声学条件的能力，在自动语音识别（ASR）中得到了广泛应用。然而，ASR 的输出在许多现实场景中仍存在质量问题，这归因于环境噪声、重叠语音、长尾或未登录词以及不同的说话人口音。

几种方法已被应用于 ASR 校正。例如，Leng 等人 [5] 引入了 FastCorrect，一个利用预训练语言模型高效纠正 ASR 错误的框架。在此基础上，Leng 等人 [6] 提出了 SoftCorrect，该方法结合了一种软决策机制以提高校正灵活性。王等人 [7] 通过增强 Transformer 模型的实体检索能力来应对特定领域的挑战，提升了其处理专业术语的能力。顾等人 [8] 开发了 DenoisingLM，一个利用去噪自动编码器改进 ASR 输出的模型。然而，这些传统方法通常表现出有限的泛化能力，难以处理多种口音或声学条件，并且在有效支持多语言方面面临重大挑战。

大型语言模型（LLMs）[9, 10, 11, 12, 13, 14]，具备深度上下文理解和生成能力，为自动语音识别后处理提供了一个有吸引力的解决方案。最近的努力展示了 LLMs 在错误纠正和转录优化方面的有效性。例如，Ma 等人 [15] 研究了使用 LLMs 对 N-best ASR 输出进行 ASR 错误纠正，采用约束解码技术和零样本应用显示出了很有前景的结果。同样地，Hu 等人 [16] 引入了一种多模态生成纠错方法 ClozeGER，该方法使用 SpeechGPT 显著提升了多个标准数据集上的转录保真度。进一步解决 ASR 幻觉问题方面，Fang 等人 [17] 提出一个三阶段基于 LLM 的框架——结合错误预检测、连贯性优化和基于推理验证，在 AISHELL 和 LibriSpeech 基准测试中显著降低了 CER/WER。此外，Sachdev 等人 [18] 探索了基于进化提示设计的后 ASR 误差纠正，提升了纠错模型的适应性和性能。

尽管在 ASR 后处理方面取得了进展，现有的方法仍然面临几个限制：1) 某些识别错误在句子上下文中看似合理，这使得大型语言模型难以检测和纠正。2) 识别错误往往是相互依赖的而非孤立的，其中一个错误可能会强化或触发其他错误，从而复杂化一步校正过程。3) 纠错模型可能生成语义上不一致但语法合理的替代词，偏离了原始语音内容。这些挑战共同强调了像我们所提出的迭代优化机制的重要性，这种机制可以在保持语义准确性的同时逐步解决相互依赖的错误。

为弥合这一差距，我们提出了 LIR-**自动语音识别**，这是一个通过 LLM 优化的启发式 ASR 校正框架。LIR-ASR 借鉴了人类听觉过程的灵感。当我们怀疑误听时，通常会尝试用音素相似的替代词来替换可能错误的声音，并评估它们在上下文中的适当性。这模拟了“听取—想象—精炼”（LIR）策略：1) 听取：解释初始 ASR 输出。2) 想象：生成不确定单词的合理音素变体。3) 精炼：在更广泛的语境中评估这些变体，以确定最准确的转录。通过引入启发式优化，“想象”阶段增加了受控随机性，使系统能够摆脱局部最优解并探索更大的解决方案空间。这种动态调整增强了 LLM 生成更准确转录的能力，尤其是在多语言环境中。在英语和中文 ASR 输出上的实验表明，LIR-ASR 相比基线平均降低了 CER/WER 高达 1.5 个百分点，显示了转录准确性方面的显著提升。总之，我们的主要贡献包括：

- 基于启发式优化的自动语音识别纠错框架，LIR-**自动语音识别**，利用大型语言模型有效处理语境上合理的识别错误。
- 一种带有有限状态机（FSM）的迭代启发式优化策略，防止校正过程陷入局部最优，并允许相互依赖的识别错误逐步得到解决。
- 基于规则的约束设计用于指导校正过程，减少由大语言模型引入的语言上合理但语义不一致的替换的风险。

2. 方法论

LIR-ASR 框架包含两个主要的架构组件：一个用于控制邻居搜索策略的有限状态机（FSM）和一个用于 ASR 校正的启发式优化模块，如图 1 所示。由 FSM 引导的邻居搜索实现了“想象”阶段，系统地指导候选空间探索策略。同时，迭代启发式优化模块支持“精炼”阶段，确保通过每次迭代逐步优化 ASR 输出。

2.1. 通过有限状态机控制搜索策略

为了更好地调节邻域搜索策略，我们引入了一个 FSM，它动态地在三种状态之间切换——无搜索、搜索和搜索++，如图 1(a) 所示。该过程以迭代计数器设置为 $i = 0$ 和连续无变化计数器设置为 $k = 0$ 初始化。然后，FSM 按如下方式进行：

无搜索状态：如果转录改善（发生变化），系统将过渡到搜索状态，重置 $k = 0$ 并增加迭代计数器（$i = i + 1$）。如果没有改进（没有任何变化），FSM 保持在非搜索状态，同时增加无变化计数器和迭代计数器（$k = k + 1$,$i = i + 1$）。

搜索状态：改进后，有限状态机移动到带有 $k = 0$ 和 $i = i + 1$ 的搜索++状态。如果没有观察到改进，它会返回到不搜索状态，重置计数器（$k = 0$，$i = i + 1$）。

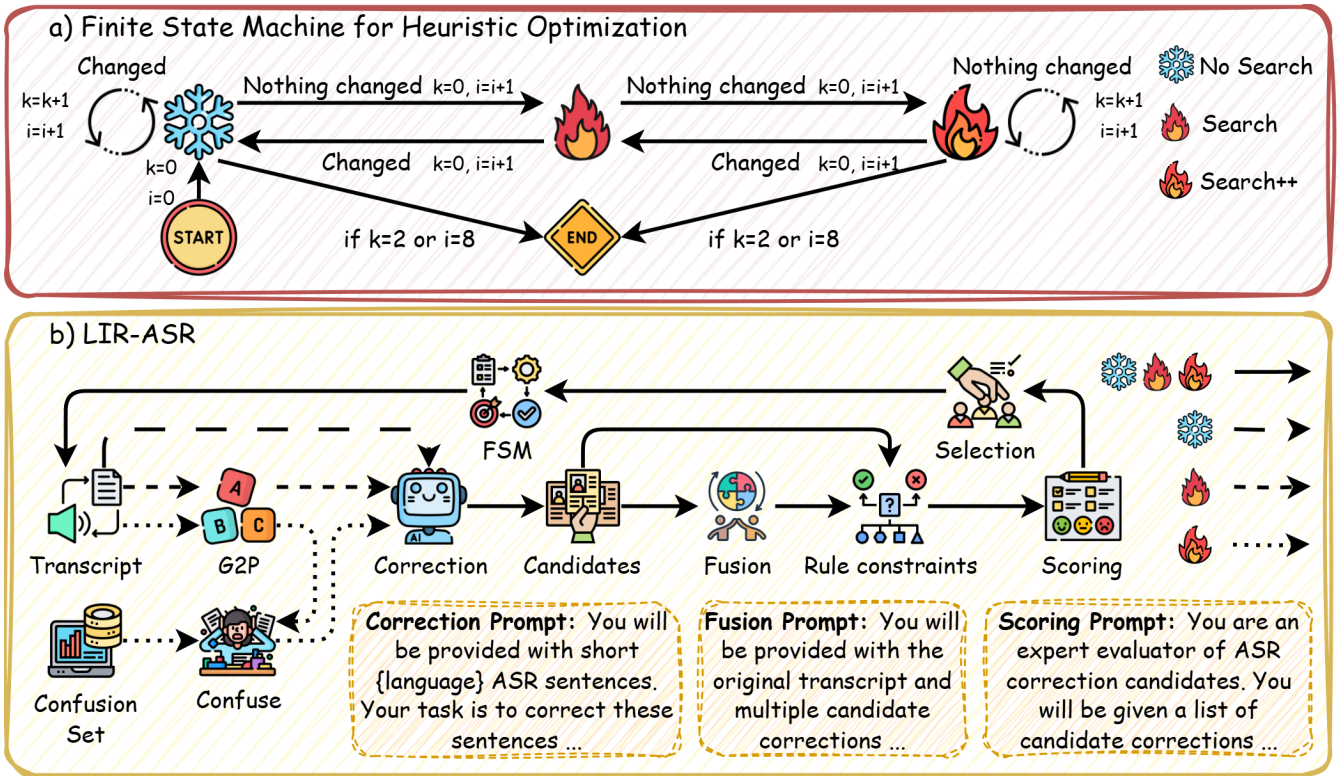


Fig. 1: 所提出的 LIR-ASR 的整体框架。

搜索++状态： 如果发生改变，有限状态机将返回到搜索状态，并使用 $k = 0$ 和 $i = i + 1$。否则，它将保持在搜索++状态，递增迭代计数器同时保持无变化计数器为零（$k = 0$, $i = i + 1$）。

终止： 优化终止，进入结束状态，如果连续无变化计数器达到预定义阈值 ($k \geq 2$) 或超过最大迭代次数 ($i \geq 8$)。

这种基于有限状态机的设计使启发式优化过程能够系统地在探索和开发之间交替，保持全局搜索与局部细化之间的平衡。终止条件防止了过多的迭代，同时避免陷入局部最优。

2.2. 语音识别修正的启发式优化

语音识别修正的启发式优化过程包含五个主要组件：邻域生成（$\mathcal{N}$）、修正（$\mathcal{C}$）、候选融合（$\mathcal{F}$）、规则约束（$\mathcal{R}$）和评分（f），共同定义了优化目标。

令转录为 s，得分为 $f(s)$，并令 $S = \{s_0, s_1, \ldots, s_n\}$ 表示从邻域 $\mathcal{N}(s)$ 生成的候选转录集合。通过音素到音标（G2P）转换和发音相似字符替换来执行邻居生成，从而产生更丰富的候选转录集。评分组件评估所有有效的候选者，并且大型语言模型为每个候选者分配一个得分和推理解释。每个候选转录都通过大型语言模型并行进行校正：

$$s'_i = \mathcal{C}(s_i), \quad \forall s_i \in \mathcal{N}(s), \tag{1}$$

其中 s'_i 表示校正后的候选转录。当存在多个候选结果时，通过额外的 LLM 步骤将它们融合成一个能够最优保持语义意义并纠正语音或打字错误的单一记录：

$$s_{\text{fused}} = \mathcal{F}(s, \{s'_0, s'_1, \ldots, s'_n\}), \tag{2}$$

其中 $\mathcal{F}(\cdot)$ 选择最符合语义一致性的候选者，整合了多份修正记录的优势。最后，当前的记录和融合后的记录被纳入候选集：

$$S' \leftarrow \{s'_0, s'_1, \ldots, s'_n, s, s_{\text{fused}}\}, \tag{3}$$

确保在后续的启发式评估中同时考虑原始结果和融合后的结果。

规则约束随后被用于过滤掉不可靠的候选词，因为基于 LLM 的校正可能会引入在语言上合理但在语义上不一致的替换。虽然邻域搜索有效增强了候选词的多样性，但不可避免地会带来额外的噪声。有两种类型的规则约束：1) **音韵一致性：** 候选词的发音必须与原词足够相似，通过语音相似度指标（如拼音或 G2P）进行衡量。**长度和结构一致性：** 过多插入或删除的候选词将被过滤掉。

令 s^t 表示第 t 次迭代的转录，$f(s)$ 是由大语言模型赋予的相应分数。在每次迭代中，贪婪接受规则会选择一个候选者 $s'_i \in S'$，其分数不低于当前转录的最大值：

$$s^{t+1} = \begin{cases} s'_i, & \text{if } f(s'_i) = \max_{s' \in S'} f(s') \geq f(s^t), \\ s^t, & \text{otherwise.} \end{cases} \tag{4}$$

由于每个步骤的候选集 S' 是有限的，并且接受规则确保

$$f(s^{t+1}) \geq f(s^t), \tag{5}$$

分数序列 $\{f(s^t)\}$ 是非递减的。此外，因为 $f(s)$ 被最大可实现得分 $f_{\max}$（例如，完美的成绩单）所上界，

$$f(s^t) \leq f_{\max}, \quad \forall t, \tag{6}$$

根据单调收敛定理，序列 $\{f(s^t)\}$ 收敛：

$$\lim_{t \to \infty} f(s^t) = f^* \leq f_{\max}. \tag{7}$$

因此，迭代优化保证收敛到一个无法在探索邻域内进一步提高得分的成绩单，即局部最优。

Table 1: 不同方法的 ASR 纠错性能对比。结果分别针对中文（ZH）和英文（EN）报告。粗体数字表示每一列中的最佳性能。* 表示提出的方法。最后一列显示了与未纠正基线相比 CER/WER 的平均改进。

大语言模型	方法	耳语介质 ZH	耳语介质 EN	Whisper-large-v3 ZH	Whisper-large-v3 EN	Δ 字错率 /Δ 句错率
none	none	5.14 / 10.21	1.92 / 4.49	3.98 / 7.50	1.75 / 4.23	–
Qwen3-235B	direct prompt	5.04 / 9.21	2.33 / 4.60	4.22 / 7.11	2.30 / 4.94	-0.28 / +0.14
	evolutionary prompt	5.11 / 8.98	2.51 / 5.09	5.64 / 9.11	2.96 / 5.54	-0.86 / -0.57
	3-best	4.80 / 9.50	1.93 / 4.35	3.94 / 7.40	1.94 / 4.38	+0.04 / +0.20
	6-best	4.54 / 9.13	1.90 / 4.36	3.88 / 7.29	1.99 / 4.57	+0.12 / +0.27
	LIR-ASR*	2.85 / 5.97	1.88 / 4.26	2.82 / 5.60	1.72 / 3.89	+0.88 / +1.68
深度搜索-V3.1	direct prompt	3.59 / 5.95	1.99 / 4.38	3.19 / 6.41	2.18 / 4.38	+0.46 / +1.33
	evolutionary prompt	4.33 / 7.53	2.22 / 4.32	3.33 / 5.46	2.20 / 4.18	+0.18 / +1.24
	3-best	3.78 / 6.62	1.93 / 4.15	3.19 / 5.36	1.98 / 4.12	+0.48 / +1.55
	6-best	3.20 / 5.95	1.60 / 3.59	2.99 / 5.34	1.91 / 3.90	+0.77 / +1.91
	LIR-ASR*	3.20 / 5.91	1.68 / 3.70	2.89 / 5.23	1.59 / 3.60	+0.86 / +2.00
	w/o rule	4.21 / 8.34	1.74 / 3.62	6.62 / 9.27	1.84 / 3.85	-0.41 / +0.34
	w/o multi-candidate	3.89 / 7.50	1.73 / 3.79	3.24 / 6.40	1.73 / 3.90	+0.55 / +1.21
	w/o FSM	3.93 / 7.56	1.71 / 3.75	3.52 / 6.46	1.69 / 3.82	+0.48 / +1.21
	w/o search	3.80 / 7.43	1.68 / 3.70	3.25 / 5.98	1.71 / 3.86	+0.59 / +1.37

3. 实验

3.1. 设置

为了评估所提出方法的校正性能，我们在 FLEURS 数据集的测试子集 [19] 上进行了实验，该数据集涵盖了 102 种语言，并包含 945 个中文样本和 647 个英文样本。为了检验其在不同语音识别系统中的鲁棒性，我们采用了 Whisper-medium 和 Whisper-large-v3 作为基础识别器。此外，为了评估其对不同校正模型的适应性，使用了 Qwen3-235B[11] 和 DeepSeek-V3.1[13]。在所有实验中，最大迭代次数设置为 8，候选池大小固定为 3。

我们的实验基线包括直接提示 [18]，进化提示 [18]，和 n 最佳 [20]。为了评估我们提出方法中各个组件的贡献，我们在基于 DeepSeek-V3.1 的 LIR-ASR 上进行了一项消融研究，系统地移除了规则约束、多候选选择、FSM 和邻居搜索。使用 CER 和 WER 作为评估指标来量化 ASR 纠正性能。

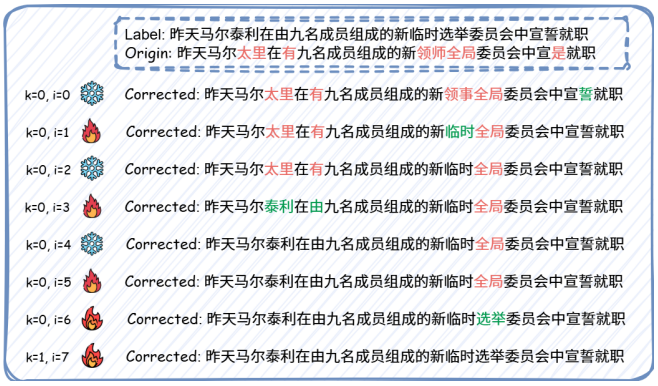


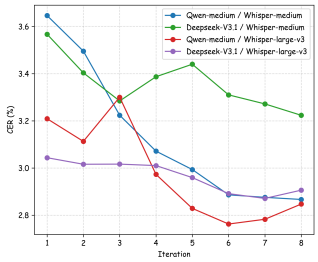
Fig. 2: ASR 校正的一个示例。

3.2. 主要结果

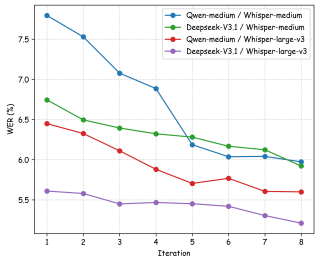
表 1 提供了对 Whisper-medium 和 Whisper-large-v3 模型在中文和英文中应用的 ASR 纠正方法的全面比较。所提出的 LIR-ASR 方法在所有设置下均实现了最低的 CER/WER。在评估的 LLM 中，DeepSeek-V3.1 与 LIR-ASR 结合时取得了最显著的改进，在 Whisper-medium 模型上中文的 CER 达到 3.20 和 WER 达到 5.91，英文的 CER 达到 1.68 和 WER 达到 3.70。这相对于基线而言平均提高了 +0.86 的 CER 和 +2.00 的 WER。值得注意的是，带有 LIR-ASR 纠正的 Whisper-medium 模型性能优于 Whisper-large-v3 基线。具体来说，对于中文，Whisper-medium 上的 LIR-ASR 实现了 3.20 的 CER，超过了 Whisper-large-v3 基线的 3.98 CER。对于英文，Whisper-medium 上的 LIR-ASR 实现了 3.70 的 WER，超过了 Whisper-large-v3 基线的 4.23 WER。LIR-ASR 在所有测试配置中表现出最高的平均改进。相对于基线的平均改进分别为 Qwen3-235B 的 +0.88/+1.68 和 Deepseek-V3.1 的 +0.86/+2.00，这表明所提出的方法在增强 ASR 纠正性能方面是有效的。虽然进化提示旨在迭代优化提示，但它缺乏明确的语义或语言约束。因此，在大多数情况下，其表现不如未纠正的基线。n-best 方法始终优于直接和进化提示方法，作为强大的基线存在。增加候选数量从 3 最佳到 6 最佳进一步提高了校正性能，正如 Qwen3-235B 在 Whisper-medium 上所见：中文的 CER 从 4.80（3 最佳）降低到了 4.54（6 最佳）。这表明扩展候选池允许系统选择更准确的替代方案。

3.3. 消融研究

消融实验结果见表 1 的最后四行。移除基于规则的约束导致性能下降最大，例如，对于 Whisper-large-v3 和 DeepSeek-V3.1 中的中文，CER/WER 从 2.89/5.23 增加到 6.62/9.27，强调了它们在引导 LLM 避免语义不一致但语言上合理的替代方案方面的重要作用。省略多候选生成导致 CER/WER 上升至 3.24/6.40，证实了生成多样化的候选变体对于有效的迭代改进和提高转录质量至关重要。排除 FSM 略微降低了性能到 3.52/6.46，表明基于状态的控制有助于启发式优化过程避免局部最优，并更好地管理相互依赖的识别错误。最后，移除邻域搜索机制导致 CER/WER 增加至 3.25/5.98，证明了在“想象”步骤中引入受控随机性可以探索更广泛的解决方案空间，支持逐渐收敛到更准确的修正。



(a) CER 与 迭代次数



(b) WER 与 迭代次数

Fig. 3: 所提 LIR-ASR 的收敛可视化

3.4. 收敛性分析

为了评估迭代优化在所提出的 LIR-ASR 框架中的有效性，图 2 展示了一个中文示例。如图所示，通过连续的迭代逐步纠正了 ASR 错误。明显错误在 No Search 和 Search 阶段得到解决，而更细微、几乎不易察觉的错误则在 Search++阶段被解决。

为了进一步评估我们方法的收敛行为，我们跟踪了每次迭代生成的转录本的 WER 和 CER 指标。图 3 展示了代表性中文样本在各次迭代中的 WER 和 CER 曲线。这些曲线表现出单调非增的趋势，表明启发式优化始终能提升转录质量。此外，在几次迭代后改进逐渐趋于平稳，证明该方法能在探索的候选邻域内收敛到一个稳定解。这些结果证实了我们的迭代启发式优化不仅有效提升了 ASR 输出，还在不同语言和 ASR 基础模型上表现出可靠且稳定的收敛行为。

4. 结论

在本文中，我们提出了 LIR-ASR，一种新颖的模拟启发式优化迭代 ASR 校正框架，利用大型语言模型。受人类听觉感知的启发，LIR-ASR 采用了一种“倾听-想象-精炼”的范式，在该范式中，首先识别不确定的词汇，然后生成语音上合理的替代词，并最终在转录文本的上下文约束内进行精炼。通过集成有限状态机来控制启发式搜索过程并应用基于规则的约束来指导 LLM 评分，我们的方法有效地减轻了语境合理性的识别错误，并减少了语义不一致替换的风险。对英语和中文 ASR 输出的实验表明，LIR-ASR 相较于基线平均减少了高达 1.5 个百分点的 CER/WER，展示了转录准确性上的显著提升。在未来的工作中，我们计划将 LIR-ASR 扩展到藏语和其他低资源语言，在这些语言中，语音变异和数据稀缺性带来了额外的挑战，进一步验证其跨语言适应能力。

致谢

本工作部分由中国国家自然科学基金资助项目 62276055 和 62406062 支持，部分由四川省科技计划项目资助编号 2023YFG0288 支持，部分由四川省自然科学基金资助项目 2024NSFSC1476 支持，部分由国家重点研发计划项目编号 2022ZD0116100 支持，部分由四川省重大科技项目，资助编号 2024ZDZX0012 支持。

5. REFERENCES

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

[2] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, and et al., “Google usm: Scaling automatic speech recognition beyond 100 languages,” 2023.

[3] Yangyang Meng, Jinpeng Li, Guodong Lin, Yu Pu, Guanbo Wang, Hu Du, and et al., “Dolphin: A large-scale automatic speech recognition model for eastern languages,” 2025.

[4] Yosuke Kashiwagi, Hayato Futami, Emiru Tsunoo, and Satoshi Asakawa, “Whale: Large-scale multilingual asr model with w2v-bert and e-branchformer with large speech data,” 2025.

[5] Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiang-Yang Li, Ed Lin, and Tie-Yan Liu, “Fastcorrect: Fast error correction with edit alignment for automatic speech recognition,” 2022.

[6] Yichong Leng, Xu Tan, Wenjie Liu, Kaitao Song, Rui Wang, Xiang-Yang Li, Tao Qin, Edward Lin, and Tie-Yan Liu, “Softcorrect: Error correction with soft detection for automatic speech recognition,” 2023.

[7] Haoyu Wang, Shuyan Dong, Yue Liu, James Logan, Ashish Agrawal, and Yang Liu, “Asr error correction with augmented transformer for entity retrieval,” 2020.

[8] Zijin Gu, Tatiana Likhomanenko, He Bai, Erik McDermott, Ronan Collobert, and Navdeep Jaitly, “Denoising lm: Pushing the limits of error correction models for speech recognition,” 2024.

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al., “The llama 3 herd of models,” 2024.

[10] Team GLM, , Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, and et al., “Chatglm: A family of large language models from glm-130b to glm-4 all tools,” 2024.

[11] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al., “Qwen3 technical report,” 2025.

[12] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and et al., “Gpt-4 technical report,” 2024.

[13] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, and et al., “Deepseek-v3 technical report,” 2025.

[14] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, and et al., “Kimi k2: Open agentic intelligence,” 2025.

[15] Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill, “Asr error correction using large language models,” 2025.

[16] Yuchen Hu, Chen Chen, Chengwei Qin, Qiushi Zhu, EngSiong Chng, and Ruizhe Li, "Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models," in Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, Aug. 2024, pp. 666–679, Association for Computational Linguistics.

[17] Yangui Fang, Baixu Cheng, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, and Guohui Zhong, "Fewer hallucinations, more verification: A three-stage llm-based framework for asr error correction," 2025.

[18] Rithik Sachdev, Zhong-Qiu Wang, and Chao-Han Huck Yang, "Evolutionary prompt design for llm-based post-asr error correction," 2024.

[19] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, and et al., "Fleurs: Few-shot learning evaluation of universal representations of speech," 2022.

[20] Rao Ma, Mark J. F. Gales, Kate M. Knill, and Mengjie Qian, "N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space," in INTERSPEECH 2023. Aug. 2023, p. 3267 – 3271, ISCA.