瑞士议会语料库重新构想 (SPC_R)：基于 RAG 的校正和预测 BLEU 的增强转录

Vincenzo Timmel[1], Manfred Vogel[1], Daniel Perruchoud[1], Reza Kakooee[1]

[1]University of Applied Sciences and Arts Northwestern Switzerland

{ vincenzo.timmel, manfred.vogel, daniel.perruchoud, reza.kakooee } @fhnw.ch

Abstract

本文提出了瑞士议会语料库的新长格式版本，将整个瑞士德语辩论会议（每个都与官方会议记录对齐）转换成高质量的语音-文本对。我们的流程首先使用 Whisper Large-v3 在高计算环境下将所有会议音频转录为标准德语。然后，我们应用两步 GPT-4o 校正过程：首先，GPT-4o 将原始 Whisper 输出与官方记录一起输入，以修正识别错误，主要是命名实体。其次，一个单独的 GPT-4o 步骤评估每个修正后的段落的语义完整性。我们过滤掉任何 Whisper 的平均标记对数概率推测 BLEU 分数和 GPT-4o 评估分数低于某一阈值的段落。最终语料库包含 801 小时的音频，其中 751 小时通过我们的质量控制。与原来的句子级别 SPC 版本相比，我们的长格式数据集实现了 6 点 BLEU 的提升，这证明了结合强大的 ASR、基于 LLM 的校正和数据驱动的过滤对于低资源、特定领域的语音语料库的重要性。

在低资源领域，数据匮乏仍然阻碍了自动语音识别（ASR）系统的发展。对于瑞士德语，(Plüss et al., 2021) 贡献了瑞士议会语料库（SPC），该语料库包括一份精心准备的训练数据集，其包含 176 小时的瑞士德语演讲，这些演讲与伯尔尼议会辩论的标准德语文本转录对齐，并附有 6 小时的经过精心策划的测试数据集。该语料库是通过一种强制句子对齐程序和对齐质量评估器构建的，克服了瑞士德语音频和标准德语文本之间的句子重排和语言不匹配等挑战。他们使用了一种基于 Needleman-Wunsch 的全局对齐算法和交集覆盖率（IoU）评估器来筛选出质量差的对齐。此外，字数每秒限制和语言检测等额外筛选确保只有准确对齐的句子被包含在内。

本文介绍的 SPC_R 语料库是原始 SPC 语料库的扩展，专注于创建、管理和发布针对瑞士德语 NLP 应用的定制数据集。最初，从伯尔尼州议会辩论中爬取的数据包括 801 小时的长格式会议录音，每场会议长度从 28 分钟到 242 分钟不等，并配有官方会议记录。

与通过在自动生成的转录和官方会议记录之间找到近乎完美的匹配从议会会议中提取句子的 (Plüss et al., 2021) 相比，我们在 SPC_R 中引入了一个先进的转录流程。这包括使用 Whisper Large-v3 模型 (Radford et al., 2023) 进行转录，以及使用 GPT-4o (Hurst et al., 2024) 进行的转录后纠正步骤，以与官方协议保持一致，从而进一步提高转录质量和整体数据准确性。

此外，SPC_R 语料库提供的是长篇形式的数据，而原始 SPC 是按句子级别进行分割的。

主要贡献包括：

- 通过 Whisper Large-v3 高计算设置对大约 801 小时音频进行高质量转录，详见第 2 节。
- 基于通过线性回归的 Whisper 转录输出的 BLEU 分数 (Papineni et al., 2002) 预测。
- 一种两步大语言模型（LLM）方法，其中第一个模型校正转录，第二个独立的模型评估该校正。

本文详细介绍了瑞士德语未来 NLP 数据集发布的方法、实验结果及影响。

1 相关工作

在过去的几年中，几个倡议 (Plüss et al., 2021, 2022, 2023; Dogan-Schönberger et al., 2021) 为瑞士德语 ASR 解决方案的发展做出了宝贵贡献；发布的数据集概览如图 1 所示。然而，这些数据集都是在句子层面的，这通常不能改善用于真实世界情境的 ASR 解决方案 (Timmel et al., 2024) 。此外，并非所有现有的数据集都能用于商业目的。

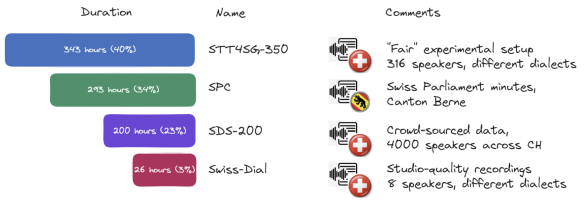

Figure 1: 瑞士德语语音到德文文本的数据集概述。SPC 可以在 MIT 许可下使用，SDS-200 和 STT4SG-350 在 SwissNLP 许可下使用。SwissDial 仅可用于研究目的。

2 使用 Whisper Large-v3 进行转录

构建 SPC_R 语料库的起点是来自伯尔尼州议会辩论的 801 小时长篇音频，我们使用 Whisper Large-v3 进行转录。

我们的转录流程在高计算设置下通过 WhisperX (Bain et al., 2023) 使用 Whisper Large-v3，具体设置为 beam_size 设为 10，best_of 设为 10，log_prob_threshold 设为-2。所有转录工作都在一台搭载 20GB VRAM 的 NVIDIA A4500 GPU 上完成，使用 float16 精度和 batch_size 为 8。这些高计算设置进一步改善了结果，如图 5 所示。对于所有转录的议会会议，我们存储 Whisper 的 avg_log_prob 输出，该输出反映了模型的预测置信度，并且在转录质量上表现出强大的预测能力，如小节 **??** 中所述。

我们观察到，Whisper 计算的置信度指标（如方程式 1 所示）与使用 Whisper 转录的数据集的 BLEU 分数（更准确地说是 sacreBLEU [1]）之间存在线性关系。

$$\text{confidence} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) \tag{1}$$

[4pt]

置信度来源于 Whisper 的段特定平均对数概率 avg_log_prob，这些概率在整个音频文件中进行平均。在公式 (1) 中，p_i 表示第 i 段的平均对数概率，N 是整个音频文件中段的总数，其中一个段是 Whisper 预测的两个时间戳之间的文本。因此，置信度是对整个音频文件的平均 avg_log_prob 的指数。

[1] https://github.com/mjpost/sacreBLEU（默认设置：4-gram，标准标记化和平滑）

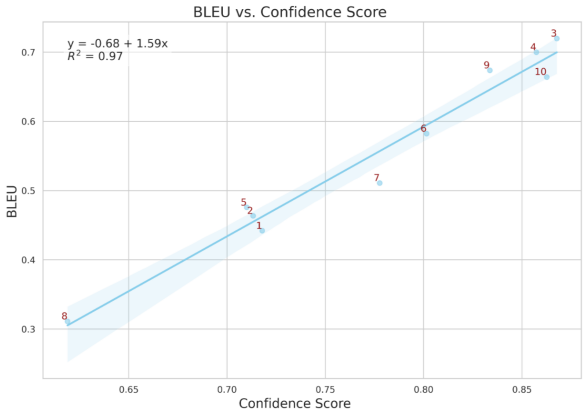

Figure 2: 对于十个长对话，用数字 1-10 表示的 BLEU 分数与 Whisper 置信分数之间的线性关系。蓝色阴影区域表示 95 % 置信区间。

图 2 展示了 BLEU 分数（在转录与人工创建的真实值之间计算得出）与十个不同且独立的瑞士德语数据集的置信度之间的线性关系。每个大约一小时（约 8,000 个标记）的数据集由两个或更多说话者之间人工转录的瑞士德语对话（真实值）构成（由于数据隐私和保密协议限制，这些数据集无法披露）。我们的分析显示，较高的置信度值与较高的 BLEU 分数呈近线性关系，这表明置信度指标是转录质量的强预测因子，暗示其在评估转录性能中的潜力。

对这些数据进行线性回归拟合得到的截距为-0.68，斜率系数为 1.59，可以仅根据置信度预测一个 BLEU 得分，称为预测 BLEU，而无需首先创建一个真实值。

图 3 显示了所有 131,291 个 SPC_R 片段的预测 BLEU 分数分布，总共对应 801 小时的音频。

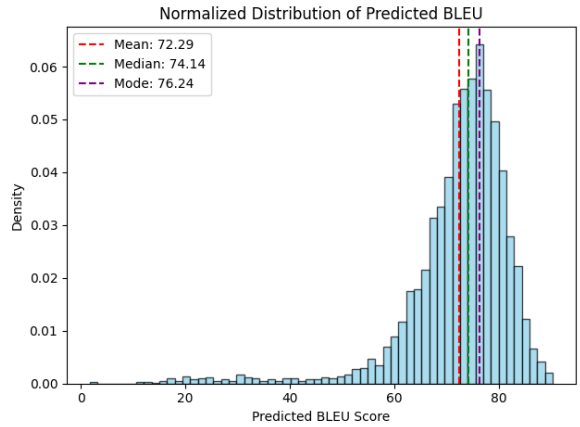

Figure 3: 预测的 BLEU 评分在 SPC_R 上的分布（N = 131’291 个数据段）。

图 4 显示了给定预测 BLEU 分数阈值的数据样本累计比例。随着阈值的提高，符合条件的

样本减少，这突显了转录质量与可用数据量之间的平衡。

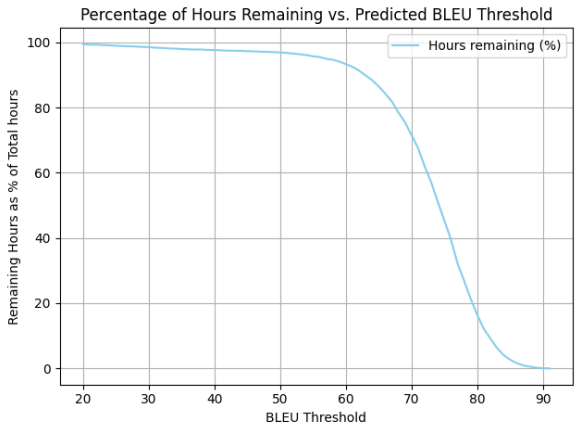

Figure 4: BLEU 分数高于阈值的数据样本百分比。

因此，可以利用从 Whisper 的 avg_log_prob 推导出的预测 BLEU 分数来识别和选择高质量的转录片段（见第 4 节）。

3 使用 GPT-4o 进行转录校正

采用 Whisper Large-v3 进行自动转录显示出良好的结果，但在命名实体（例如，将“Alberucci”转成“Alba Rutschi”）和其他类似错误方面会导致问题。为了解决这些问题，我们引入了一个使用 text-embedding-3-large GPT-4o 和 GPT-4o-mini (OpenAI, 2023) 进行两步纠正过程：

1. 修正阶段：使用 GPT-4o 逐段提示纠正错误，以此来完善初始转录。纠正内容是基于从官方会议记录概要中提取的信息，这些概要对应于音频片段，并使用检索增强生成（RAG，见子节 3.1 ）。

2. 评估阶段：对 GPT-4o 的修正进行评估时，使用手动检查小数据样本和 GPT-4o-mini-as-a-Judge。

GPT-4.1 (OpenAI, 2025) 也进行了评估，但我们发现它会反复改变单词的变位形式，因此有时在转录中引入新的错误。虽然总体上仍减少了 WER，但它纠正的错误比 GPT-4o 少。

3.1 通过 RAG 提供上下文

RAG (Lewis et al., 2020) 用于为 GPT-4o 提供事实背景以纠正转录。

我们遵循最佳实践 (Wang et al., 2024)，使用 Faiss (Douze et al., 2024) 进行高效的向量存储和检索，采用滑动窗口方法和 text-embedding-3-large 作为嵌入模型。官方手册摘要使用 pyPDF (Fenniak et al., 2024) 以 600 字符为一块并有 450 字符重叠的方式进行摄取。这些值的选择是为了在转录和来自块的上下文之间根据最大片段长度为 423 字符的设置始终确保完整的重叠。我们将最相关的块作为上下文传递给 GPT-4o，而不对检索到的块重新排序。

对 122 段音频片段进行手动评估，这些片段对应于 50 分钟的转录数据，结果显示，官方手册摘要中的正确片段被检索到了 94.1 % 的片段。这一高比率可能是由于会话协议与会话转录的对齐容易实现。

3.2 校正阶段

在校正阶段，GPT-4o 被提供了来自子节 3.1 的上下文以及需要校正的抄本，并附有一个广泛的、迭代扩展的系统提示，指定使用检索到的块和与伯尔尼方言特性相关的附加规则 [2] 。

在高计算设置下运行的流水线在使用 50 分钟的人工转录数据评估时，将词错误率（WER）从 15.7 % 改善到 11.1 %，温度设置为 0.1 以减少变异性并降低 WER（见图 5 ）。

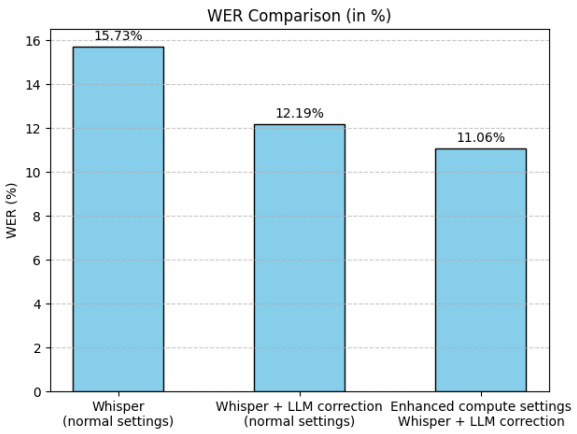

Figure 5: Whisper Large-v3 在三种配置下的词错误率（WER）：标准设置，应用 GPT-4o 校正后的结果，以及使用高计算设置（增强设置）结合 GPT-4o 校正的情况。

此外，在手动检查地点、名称、法律引用和政党等命名实体时，命名实体转录的正确性从最初使用 Whisper Large-v3 的 72.2 %（72 个中的 52 个）提高到在应用 GPT-4o 校正后的 100 %（72 个中的 72 个）。

表 1 显示了音频示例、初始 Whisper Large-v3 转录、检索的上下文以及使用 GPT-4o 校正后的输出。

[2] 规则包括诸如“vo dr”（音频）需要从“vor der”更正为“von der”，以及“mier”（音频）需要从“mir”更正为“wir”的情况。

Table 1: 示例音频输入，使用 Whisper Large-v3 进行初始转录，提供给 GPT-4o 的检索上下文（已缩短），以及其输出。鼓励 GPT-4o 尽量保持校正与输入内容接近，以便数据仍可用于训练依赖于对齐音频和文本的 ASR 系统。

Audio Input (transcribed) dass ehr au verdaut händ, wenn ehr näbem outo send.
Whisper Large-v3 output (initial transcription) dass er auch verdauert hat, wenn er neben dem Auto sitzt.
Context retrieved via RAG (given to GPT-4o as help for the correction.) sodass Sie wieder leicht ernüchtert sind und verdaut haben, wenn Sie beim Auto ankommen werden.
GPT-4o output (final, corrected transcription) dass Sie auch verdaut haben, wenn Sie neben dem Auto sind.

3.3 评估阶段

在此阶段，转录的质量在以下类别中进行评估（以下简称为判断标记）：

- 3) 完全正确：所有名称、名词、数字和缩写都被准确地转录，没有任何错误。
- 2) 小错误（不影响关键术语）：所有名称、名词、数字和缩写都是正确的。存在小的语法错误（例如，不正确的动词变化或冠词）。
- 1) 关键术语错误：在转录中至少有一个名称、名词、数字或缩写不正确。
- 0) 没有相关摘录：提供的摘录不包含任何相关内容，使得评估和纠正变得不可能。

图 6 显示了评估阶段的输出：78.0 % 的转录文本在语义上是相同的，这意味着经过 GPT-4o 的修正后，内容在转录中得到了完美反映。

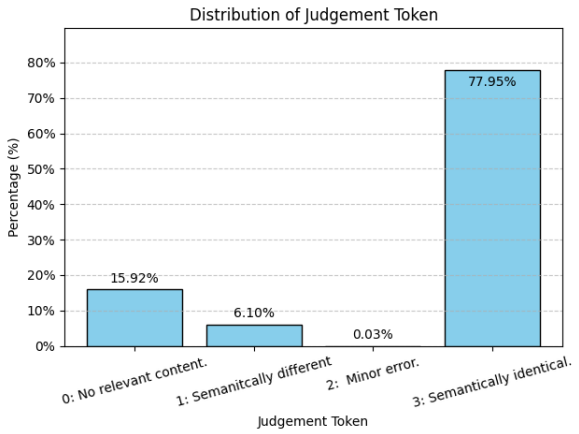

Figure 6: 使用 GPT-4o-mini-as-a-judge 对最终转录质量进行分类的分布。

经过对 50 分钟的数据分析，我们发现只有将标签 “token 0” 合并为 “token 1”，并将 “token 2” 与 “token 3” 合并时，判断类别才是可靠的。这样分组可以将分类准确率提高到 92.2 %。由于 GPT-4o-mini 在判断错误是因为缺失上下文还是由于转录中的语义发生实际变化时有困难，我们在最终的数据选择中将这些标签进行合并。

4 选择数据和训练/测试划分

为了构建高质量的 SPC_R 语料库，我们结合来自第 **??** 节（预测 BLEU）和第 **??** 节（判断标记符）的发现，如图 7 所示。

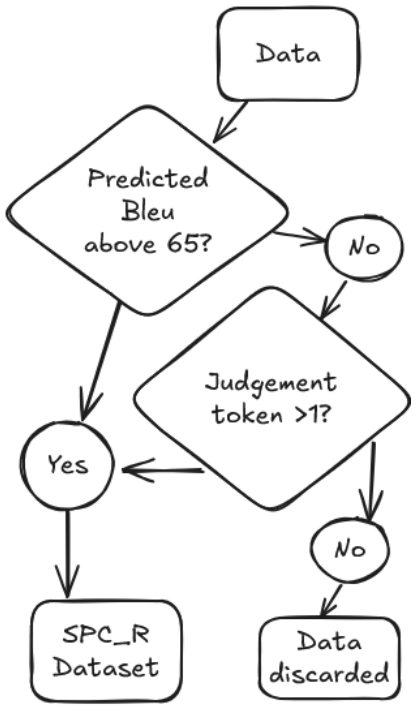

Figure 7: 用于构建高质量 SPC_R 语料库数据集的逻辑。初始数据集 “Data” 的大小为 801 小时音频，高质量数据集 “SPC_R” 的大小为 751 小时。

基于之前的研究 (Cloud)，我们选择了一个 65 的预测 BLEU 分数阈值进行筛选，这表明 BLEU 分数高于 60 可以表示转录质量优于一般人的水平。通过选择略高的阈值，我们减少了图 2 中由 95 % 置信区间所指出的变异性。虽然这不能保证数据的完美，但 (Timmel et al., 2024) 显示了当与高质量训练数据结合使用时，不完美的、伪标记的数据可以提高 ASR 模型的质量。

这导致了一个高质量的语料库，包括 751 小时的瑞士德语音频以及相配对的标准德语转录。对于测试集，选择了 50 个小时，其蓝色得分至少为 70，并且段落被评估为类别 3（如 3.3 节中所述）。因此，训练/测试的划分是 701/50 小时。

5 可用性和许可

数据集可以在 Hugging Face 上的 i4ds/spc_r 找到，完整代码库（包括提示）可以在 GitHub 上的 i4ds/spc_r 公开获取。

此数据集是根据知识共享署名 4.0 国际（CC BY 4.0）许可证发布的，该许可证允许在给予适当信用的情况下共享和改编，并且任何衍生作品都需在相同的条款下获得许可。[3]

我们展示了 SPC_R，它使用 Whisper Large-v3 在高计算设置下转录，通过 GPT-4o 在语境中进行校正，并通过 GPT-4o-mini 评估质量。这个过程产生了一个包含 751 小时高质量的瑞士德语口语与标准德语文本对应的语料库。

6 未来工作

为进一步增强瑞士议会语料库，有几个很有前景的途径。比如，除了伯尔尼议会辩论之外，加入额外的数据来源可以扩大数据集的辩证和上下文的多样性，从而可能提高瑞士德语语音识别模型的性能和鲁棒性。探索替代的转录模型，特别是开源解决方案，可能会在成本或性能上比基于 OpenAI 模型的当前方法更具优势。最后，还有机会采用更精细的评估指标，如 Para $_{both}$ (Paonessa et al., 2023)，其能够更好地捕捉语义忠实度和命名实体的准确转录。

7 局限性

评价指标：我们的评估主要依赖于标准指标，如 BLEU 和 WER。这些指标虽然有用，但并不能捕捉转录质量的所有方面，因为如果一个句子用不同的词传达正确的语义时，这些指标可能会产生误导，特别是在正确转录命名实体方面，因为它们没有权衡命名实体错误对转录理解的更大影响。根据我们的经验，Whisper 的大多数错误，至少在瑞士德语中，会在命名实体上，这降低了转录的理解。

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Google Cloud. Evaluate models | cloud translation. https://cloud.google.com/translate/docs/advanced/automl-evaluate. Accessed: 2025-03-12.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *Preprint*, arXiv:2103.11401.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and pypdf Contributors. 2024. The pypdf library.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

OpenAI. 2023. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2025-02-28.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Claudio Paonessa, Dominik Frefel, and Manfred Vogel. 2023. Improving metrics for speech translation. *arXiv preprint arXiv:2305.12918*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Christian Scheller, Yanick Schraner, Claudio Paonessa, Larissa Schmidt, Julia Hartmann, Tanja Samardzic, Manfred Vogel, and Mark Cieliebak. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. In preparation.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an

[3] 更多详情请参见 https://creativecommons.org/licenses/by/4.0/ 。

automatically aligned swiss german speech to standard german text corpus. In *Proceedings of the Swiss Text Analytics Conference*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Vincenzo Timmel, Claudio Paonessa, Reza Kakooee, Manfred Vogel, and Daniel Perruchoud. 2024. Fine-tuning whisper on low-resource languages for real-world applications. *arXiv preprint arXiv:2412.15726*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.