信息检索方法在协调社会科学中的纵向调查问题方面表现良好吗？

Wing Yan Li*
University of Surrey
United Kingdom
wingyan.li@surrey.ac.uk

Zeqiang Wang
University of Surrey
United Kingdom
zeqiang.wang@surrey.ac.uk

Jon Johnson
University College London
United Kingdom
jon.johnson@ucl.ac.uk

Suparna De
University of Surrey
United Kingdom
s.de@surrey.ac.uk

Abstract

在长期社会科学调查中，自动检测语义等价的问题对于为社会、经济和健康科学的实证研究提供信息的长期研究至关重要。检索等价问题面临着双重挑战：在不同研究之间以及在问题和回答选项之间，理论构念（即概念/子概念）的表示不一致，以及纵向文本中词汇和结构的演变。为了应对这些挑战，我们由计算机科学家和调查专家组成的多学科合作团队提出了一项新的信息检索（IR）任务，即识别问题和回答选项之间的概念（例如住房、工作等）等价性，以协调长期人口研究。本文研究了多个无监督方法在涵盖 1946-2020 年的调查数据集上的应用，包括概率模型、语言模型的线性探测以及专门用于 IR 的预训练神经网络。我们展示了专用于 IR 的神经模型达到整体最高性能，而其他方法表现相当。此外，用神经模型对概率模型的结果进行重新排序仅引入了最大为 0.07 的 F1 分数适度提升。调查专家的定性事后评估显示，模型对词汇重叠度高的问题普遍敏感性低，尤其是在子概念不匹配的情况下。总之，我们的分析进一步推动了社会科学中协调纵向研究的研究。

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Unsupervised learning*; **Ensemble methods**; **Neural networks**.

Keywords

Natural Language Processing, Information Retrieval, Conceptual Comparison, Longitudinal Study

ACM Reference Format:
Wing Yan Li, Zeqiang Wang, Jon Johnson, and Suparna De. 2025. 信息检索方法在协调社会科学中的纵向调查问题方面表现良好吗？. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3730164

1 介绍

大规模的纵向社会调查对于理解跨代的社会变化至关重要。通过针对特定人群和时间背景量身定制的问题，调查捕捉到理论结构，也被称为概念或子概念（例如住房条件）。由于社会背景和目标受访者特征（如教育水平）的变化，问卷结构、语义框架和概念的使用也发生了变化，这意味着调查在语言、结构和回应选项上各不相同。这对协调跨不同时间段和人群的研究构成了重大挑战。

Table 1: 带有相应 概念 和/或 子概念 标记的调查问题类型。加粗文本为问题，斜体文本为回答选项。

Question Type	Example
Standard	What is his 目前工作? 1. 自雇人士 | 2. 经理 | 3. 领班 4. 员工
Qualified	The thought of 伤害自己 has occurred to me: 1. yes, quite often | 2. sometimes | 3. hardly ever | 4. never
Compound	What is she 过敏的 to? (tick all that apply) 狗 1. yes

在这项工作中，我们解决并研究了在纵向调查中检测概念上等价的问题时的两个主要挑战：(1) 尽管在词汇和句法上有表面差异，如何捕捉语义等价性所带来的语言挑战。例如，“住房”可能随着时间演变为“家庭”，但其底层概念保持不变。(2) 等价概念可能在问题文本和回答选项之间以不同方式分布的结构挑战。在表格 1 中，标准问题的回答类别中的子概念可能会从根本上改变被测量的概念。相反，合格的问题不会捕捉任何子概念。

尽管信息检索（IR）技术近年来取得了显著进展，其在调查协调中的应用仍未被充分探索。传统的方法如 BM25 [14] 在词汇匹配方面表现出色，但可能忽略了由于术语变化而被隐藏的语义等价。当代神经模型 [7, 8, 11, 12] 虽然在语义建模上很强大，但容易过度拟合于表面的模式而未能捕获调查概念的领域特定性质。此外，现有的工作很少解决结构化调查问卷的独特需求，在这种问卷中，问题文本和编码的回应选项共同定义了测量构造。在计算机科学家和调查专家之间的合作中，我们通过三项主要贡献来弥合这一差距：

(1) 我们将纵向调查中检测概念等价性的问题表述为信息检索问题，特别是解决将问题和回答文本整合为连贯的输入序列的挑战。
(2) 我们通过比较概率模型（BM25）、线性探测语言模型（DeBERTa, Qwen）以及专用于信息检索的神经网络（M3-Embedding），对跨越 75 年（1946-2020）的数据集上的无监督信息检索方法进行了系统评价。
(3) 我们结合基于主题代码对齐的定量指标和领域专家的定性分析，以揭示当前信息检索方法在调查协调中的潜力和局限性。

我们的结果表明，尽管针对信息检索优化的神经模型取得了整体最高性能（$F1 = 0.79$），传统方法如 BM25 仍然具有竞争力（$F1 = 0.75$）。值得注意的是，定性分析揭示所有模型

*Corresponding author

SIGIR '25, Padua, Italy.

ACM ISBN 979-8-4007-1592-1/25/07
https://doi.org/10.1145/3726302.3730164

在处理具备高词汇重叠但子概念不同的问题时都遇到了困难，这表明需要更复杂的方法来处理概念粒度的细致化。

2 相关工作

计算社会科学. 调查数据协调是指通过研究、概念和时间等变量使不同的调查更加可比和一致的过程。现有的方法严重依赖于由专家手动注释的分类方案。这些方法通常是基于统计或程序的 [3]，因此在进行跨文化研究时会遇到挑战（例如语言障碍）[17]。另一方面，[16] 提出使用社会科学文献训练领域特定的 Word2Vec 嵌入。他们对各种任务（如使用从社会学词典收集的数据进行文本分类）的评估显示出很有希望的结果。

传统的 IR 系统继续采用基于统计证据的概率方法（例如，词频 [14, 15]）。最近，神经 IR 模型通过基于潜在语义信息执行搜索而获得越来越多的关注。预训练语言模型如 DeBERTa [6] 通常被用作阅读器模型，用于将序列嵌入到下游检索模型中 [8, 11, 12]。一些研究也提出了各种训练策略 [4, 9] 和结构设计 [1, 18] 的 IR 专用模型，旨在结合句法和语义特征的优势。

3 实验

假设有 n 组问卷 $S = \{S_1, S_2, ..., S_n\}$。令 $U = \bigcup_{i=1}^{n} S_i$ 为所有问卷的并集。对于每个查询问题及其对应的响应选项 $(q, r) \in U$，实验的目标是识别 $U - \{(q, r)\}$ 中所有潜在的概念等价。

数据集. 我们使用来自一个大型纵向社会科学调查项目的问题集，该项目由 318 份问卷组成。这些问题涵盖了从 1946 年到 2020 年不等的多样纵向跨度。总共有 42,161 个问题，采用三种不同的类型: 标准问题、合格问题和复合问题，如表 1 所示。标准问题通过在问题和回答选项中分别展示概念和子概念，具有最为多样化的概念分布。这带来了固有的复杂性和额外的挑战，因为标准问题仅以代码列表问题的形式出现。因此，我们考虑只使用代码列表问题[1]，并构建输入序列为问题和回答选项的连接。

在接下来的实验中，所有模型都无需训练。我们的基线模型是 BM25[2]，这是一个基于词频概率模型，用于检索等效项。我们采用对预训练的大型语言模型进行线性探测作为阅读器模型，以嵌入输入序列。遵循双编码器范式，输入（即问题）是独立编码的。考虑了各种编码器模型，包括 SBERT[3] [13] 句子转换器、DeBERTa-v3 [6]，以及像 Qwen[4] [19] 这样的预训练解码器模型。为了提高效率，我们遵循晚期交互框架 [8, 12]，其中表示用余弦距离（$1 - cosine\_similarity$）作为相似度度量进行离线预计算。

此外，我们研究了在 DeBERTa 和 Qwen 模型中生成序列表示的不同方法。(1) 平均池化序列表示（mean-rep）。我们取最终编码器或解码器层中所有标记表示的平均值。(2) 序列摘要标记（SST-rep）。对于编码器模型，我们使用 [CLS] 表示。对于解码器模型，由于解码器模型的自回归性质，我们使用最后一个标记表示。选择的 SBERT 模型使用 [CLS] 池化进行序列表示。由于该模型是端到端地使用 [CLS] 池化进行训练的，我们不研究使用 mean-rep 的影响，因为这将导致次优的性能。

对于 IR 专用模型，我们使用 M3-Embedding [1]，这是一种设计用于 IR 任务的多语言、多粒度和多功能性嵌入模型。该模型通过自我知识蒸馏进行训练，以在单一架构中协调密集、稀疏（词汇）和多向量检索信号。其基础模型为预训练的 XLM-RoBERTa [2]。在检索过程中，它使用 mean-rep 和 SST-rep 来捕捉词序和上下文细微差别。

具体来说，我们比较了两种 M3-Embedding 的变体，`BGE-m3` 和 `BGE-reranker-m3`。前者遵循双编码器范式，而后者遵循交叉编码器范式，其中查询问题和相应的候选问题成对编码。`BGE-m3` 用于端到端排序和重新排序任务，而 `BGE-reranker-m3` 仅用于重新排序任务。在两种模型中，评分函数都是上述三种检索评分的加权[5]和。

最后，对于再排序任务，神经模型被要求从 BM25 中细化每个查询的前 50 个问题。除了 `BGE-reranker-m3` 外，所有其他模型均遵循上述 LLM 线性探测设置。

评估. 问卷问题使用一个包含 16 个顶级主题（例如：教育、健康）和 120 个细化子主题（例如：教育-小学教育、教育-高等教育）的分层标签本体进行标记。我们通过比较查询问题与模型识别出的最等价匹配问题之间的顶级主题代码来评估模型性能。这基于这样一个假设：等价的问题涉及相同的一般主题。请注意，实验中没有随机性，因为我们的模型是无训练的（即：具有固定权重）。

XMONTHX_0

端到端排序比重排序更具挑战性，因为要搜索的问题更多。表格 2 展示了通过比较查询问题的顶级主题代码和对应的最相似问题来评估的模型表现。有趣的是，`BGE-m3` 在所有方面都达到了最高的性能指标，平均在所有指标上比 `BM25` 高出 0.04。可能，这得益于 `BGE-m3` 的综合评分函数，该函数从多个方面对问题进行评分。通常，F1 分数和准确率之间有 0.08 的平均差距，特别是在 `Qwen-last` 中这一差距更加明显。这一持续的

Table 2: 通过匹配查询问题和由每个模型返回的最相似问题（即 top-1）的顶层主题代码来评估模型性能。BM25 是基准。对于 DeBERTa 和 Qwen，`mean`、`[CLS]` 和 `last` 表示使用的序列表示类型。

Model	Setup	Precision	Recall	F1	Accuracy
End-to-End Ranking					
BM25		0.75	0.75	0.75	0.80
SBERT		0.73	0.55	0.62	0.66
BGE-m3		0.80	0.79	0.79	0.84
DeBERTa	mean	0.69	0.69	0.69	0.78
	[CLS]	0.68	0.68	0.68	0.77
Qwen	mean	0.75	0.72	0.73	0.83
	last	0.68	0.67	0.67	0.80
Re-Ranking					
SBERT		0.74	0.55	0.62	0.66
BGE-m3		0.77	0.78	0.77	0.83
BGE-reranker-m3		0.74	0.73	0.74	0.81
DeBERTa	mean	0.70	0.71	0.70	0.81
	[CLS]	0.76	0.74	0.75	0.80
Qwen	mean	0.74	0.71	0.72	0.83
	last	0.70	0.68	0.69	0.81

[1] 总共有 30,863 个代码列表问题。
[2] 我们使用 bm25s 库 [10] 进行高效实现
[3] 选择的基础模型是 `multi-qa-mpnet-base-dot-v1`。
[4] 我们使用具有 3B 参数和 4 位量化的 Qwen-2.5。
[5] 我们遵循了在 [1] 中建议的超参数设置。

Table 3: 标签分布以百分比形式呈现（%）。总共随机选择了 203 个问题。标签为：1 — 精确匹配；1a — 等价；2 — 子概念不匹配，3 — 完全不匹配。

Model	1	1a	2	3
BM25	65.02 %	9.85 %	13.30 %	11.82 %
BGE-m3	67.98 %	7.88 %	13.79 %	10.34 %
DeBERTa-mean	57.64 %	7.39 %	15.27 %	19.70 %
Qwen-mean	66.50 %	4.93 %	14.29 %	14.29 %

差距表明，模型在识别正面案例（即相关项）方面相对较弱，但在识别非相关项方面表现良好。

表征的类型. 使用平均表示的模型在各个方面的表现都优于使用 SST 表示的模型。从直观上来看，平均化表示将导致信息的丢失，因为该操作平滑了嵌入在表示中的分布信息。这忽略了单词顺序和单词之间交互等信息 [5] 。因此，平均表示包含的语义信息比 SST 表示要浅。可能地，这意味着嵌入相对较浅语义信息的输入序列可能对该任务更有利。

这一点也反映在重新排序的结果中。使用神经模型对 BM25 输出进行重新排序预计可以通过补偿 BM25 中语义信息的缺乏来提高性能。然而，结果显示与基线和端到端方法相比几乎没有改善或改善效果甚微，这提示在这个任务中，似乎在句法和语义信息之间取得平衡是至关重要的。

假设具有相同主题代码的问题是等价的，但这可能并不总是准确的，因为概念或子概念更加细化且具有特定问题性。因此，随机选择 203 对问题并由调查专家手动检查。通过检查查询问题及其对应的最相似问题，定义了四个标签来指示这对问题是否为：1 —完全匹配；1a —等价；2 —子概念不匹配，或 3 —完全不匹配。

在这里，我们专注于基线和端到端排序模型，特别是那些在 F1 得分和准确率上表现较好的模型变体，如 BGE-m3 和 DeBERTa-mean 。表格 3 显示了所考虑模型的标签分布。虽然这些模型在识别主要是完全匹配的情况上表现良好，类别 2 和类别 3 实例在分布中代表了其次的高频，而类别 1a 始终是最小的组。

所有模型中出现完全匹配案例的比例较高，这可能是由于识别此类案例的复杂性相对较低。此外，在不同问卷中重复出现相同的问题，例如表 4 (d)，进一步导致了第 1 类实例的普遍性。通常模型识别第 2 类实例多于第 3 类实例。这表明模型对改变潜在意义的微妙词汇变化不敏感，因此不敏感于理论结构。如表 4 (b) 所示，标记为次概念不匹配的问题对通常在词汇上有很高的重叠，但句法上略有差异。

类 1a 实例很难识别，因为它们通常涉及对查询的改写形式的问题（表 4 (a) 和 (c)）。虽然人们可能会认为需要语义理解来检测此类情况，但我们将问题和响应选项连接在一起，引入了响应列表中高度的词汇重叠。这一特性可能有利于识别最多类 1a 实例的 BM25 ，与其他模型相比。

总而言之，与之前的论述一致，在该任务中，句法和语义信息之间的平衡至关重要。识别概念上等价的问题不仅仅是找到相似的文本。模型不仅需要“理解”潜在的含义，还必须对可能改变问题概念或子概念的细微变化保持敏感。这一细微差别强调了分别调整对表层模式和更深入语义方面的重视的重要性。

4 结论

本文对社会科学研究中协调纵向调查问卷的无监督信息检索方法进行了系统评价。我们发现，尽管像 BGE-m3 这样的专用 IR 模型实现了最高的整体性能 ($F1 = 0.79$)，传统方法如 BM25 仍然出乎意料地具有竞争力 ($F1 = 0.75$)。这表明，即便语义理解有所改善，词汇匹配仍在调查协调中起着至关重要的作用。我们的定性分析进一步揭示，模型通常擅长识别准确（或等价）的匹配，但在处理细微的概念变化时存在困难，尤其是当子概念尽管具有较高的词汇重叠却有所不同时。其次，我们不同表示方式的实验表明，平衡句法和语义信息至关重要。mean-rep 优于 SST-rep 的表现表明，浅层语义建模可能更适合问卷问题匹配。这个发现挑战了常见的假设，即更深刻的语义理解总是能在文本匹配任务中带来更好的表现。最后，我们未来的工作包括：(1) 领域适应的超参数优化。对于 BGE-m3 ，系统的超参数搜索可能重新校准词汇和语义匹配信号在调查协调中的平衡。(2) 细粒度概念建模。层次化建模方法可能更好地捕捉调查概念及其组成部分之间的细致关系。(3) 特定领域语言模型。开发专门针对社会科学调查语料库进行预训练的语言模型可能更好地捕捉调查问题的领域特定术语和结构。

References

[1] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand. doi:10.18653/v1/2024.findings-acl.137

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747

[3] Joshua Kjerulf Dubrow and Irina Tomescu-Dubrow. 2016. The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. *Quality & Quantity* (2016).

[4] Jiafeng Guo, Yinqiong Cai, Keping Bi, Yixing Fan, Wei Chen, Ruqing Zhang, and Xueqi Cheng. 2025. CAME: Competitively Learning a Mixture-of-Experts Model for First-stage Retrieval. *ACM Trans. Inf. Syst.* (2025).

[5] Prakhar Gupta and Martin Jaggi. 2021. Obtaining Better Static Word Embeddings Using Contextual Embedding Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5241–5253. doi:10.18653/v1/2021.acl-long.408

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*. https://openreview.net/forum?id=XPZIaotutsD

[7] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[8] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[9] Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503* (2023).

[10] Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *arXiv preprint arXiv:2407.03618* (2024).

[11] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized Embeddings for Document Ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[12] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer Learning Approaches

Table 4: 从 203 个随机样本中选择的定性示例。加粗文本是问题，斜体文本是回应选项。"标签"分别表示检索到的最相似问题与查询问题之间的关系类型。

Index	Query Question	BM25	BGE-m3	Qwen-mean
(a)	(NCDS Age 33 Cohort Member Interview) And how satisfied or dissatisfied are you with the area you live in? 1, very satisfied | 2, fairly satisfied | ... | 8, don't know	(Wave 3 Questionnaire) On the whole, are you very satisfied, fairly satisfied, a little dissatisfied or very dissatisfied with the way democracy works in this country? 1, very satisfied | ... | 4, very dissatisfied	(1989 Main Questionnaire) How do you feel about living in this district? Would you say that you are 7, very satisfied | ... | 1, very dissatisfied	(1989 Main Questionnaire) And how do you feel about your present accommodation? 7, very satisfied | ... | 1, very dissatisfied
	Label	3	1a	1a
(b)	(Me and My Baby) What arrangements have you made about looking after your baby when you begin work? day nursery 1, yes | 2, no | 9, don't know	(Me and My Baby) What arrangements have you made about looking after your baby when you begin work? baby's grandparent 1, yes | 2, no | 9, don't know	(Me and My Baby) What arrangements have you made about looking after your baby when you begin work? other (please describe) 1, yes | 2, no | 9, don't know	(Me and My Baby) What arrangements have you made about looking after your baby when you begin work? partner 1, yes | 2, no | 9, don't know
	Label	2	2	2
(c)	(Me and My School) Do you like answering questions in class? 1, never | 2, sometimes | 3, often (most days) | 4, always	(MCS Age 7 Cohort Member Paper Self-Completion) How much do you like answering questions in class? 1, i like it a lot | 2, i like it a bit | 3, i don't like it	(Me and My School) Do you feel happy at school? 1, never | 2, sometimes | 3, often | 4, always	(Me and My School) Do you get homework? 1, never | 2, sometimes | 3, often | 4, always
	Label	1a	3	3
(d)	(Your Health Events and Feelings) Do you use gas for cooking? 1, yes, ring(s) only | ... | 4, no, not at all	(Partner and Home) Do you use gas for cooking? 1, yes, ring(s) only | ... | 4, no, not at all	(Looking After the Baby) Do you use gas for cooking? 1, yes, ring(s) only | ... | 4, no, not at all	(Partner and Home) Do you use gas for cooking? 1, yes, ring(s) only | ... | 4, no, not at all
	Label	1	1	1

for Building Cross-Language Dense Retrieval Models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I.*

[13] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[14] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* (2009).

[15] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[16] Fabian Stöhr. 2024. Advancing language models through domain knowledge integration: a comprehensive approach to training, evaluation, and optimization of social scientific neural word embeddings. *Journal of Computational Social Science* (2024).

[17] Tuan-I Tsai, Lauretta Luck, Diana Jefferies, and Lesley Wilkes. 2024. Challenges in adapting a survey: ensuring cross-cultural equivalence. *Nurse researcher* (2024).

[18] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[19] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115* (2024).