自验证改进了少量临床信息提取

Zelalem Gero [*][1] Chanda n Singh [*][1]郝[1] Cheng [1]特里斯坦[1]瑙曼[1]
Michel Galley 高剑峰潘海峰

抽象的

从非结构化数据中提取患者信息文本是健康决策支持的一项关键任务和临床研究。大型语言模型 I(LM）已显示出加速发展的潜力通过少量情境学习进行临床管理，与监督学习相反，监督学习需要昂贵的人工注释。然而，尽管有严重的尽管GPT-4 等现代法学硕士仍然 取得了进步与准确性和交互性问题作斗争美观性，尤其是在安全关键领域比如健康。我们探索一种通用的缓解措施使用自我验证的框架，该框架法学硕士的年龄为自己提供出处提取并检查其自身的输出。这个框架——工作是通过之间的不对称性而成为可能的验证和生成，其中前者是往往比后者容易得多。实验结果表明，我们的方法始终如一地证明各种法学硕士跨标准的准确性临床信息提取任务。添加盟友，自我验证产生的解释对应于每个的短文本范围的形式输出，这使得它对人类来说 是有效的有权审核结果，为可靠地提取临床信息资源受限的场景。为了方便fu在这个方向上进行真正的研究，我们发布我们的代码和提示。[1]

1。简介及相关工作

临床信息提取在临床信息提取中发挥着关键作用分析医疗记录并实现医疗保健实践操作者有效地访问和利用患者数据（Zweigen-鲍姆等人，2007年；王等人。，2018）。少样本学习方法已成为解决问题的有希望的解决方案临床信息中标记训练数据的稀缺提取任务（Agrawal 等人，2022；Laursen 等人，2023）。然而，这些方法在准确性方面仍然存在困难和可解释性，这都是医学领域 的关键问题域（Gutie´rrez 等人，2022）。

在这里，我们通过使用自我验证（SV）来解决这些问题改进少样本临床信息提取。SV 建立在最近将大语言链接在一起的作品的基础上模型（LLM）呼吁提高LLM的表现（吴等人。，2022；王等人。，2022；蔡斯，2023）。直观地说这些连锁店的成功是因为法学硕士可能能够形成任务中的各个步骤，例如。证据验证，比法学硕士更准确地执行整个任务，例如。信息提取（Ma et al.，2023；Madaan et al.，2023；张等人。，2023）。这样的连锁店已经成功了 —— 丰富的设置，例如多跳问答（按等人。，2022），检索增强/工具增强问题回答（Peng 等人，2023；Paranjape 等人，2023；Schick 等人。，2023；高等人。，2023），以及代码执行（Jojic 等人。，2023）。在这里，我们分析是否建立这样一个链可以提高临床信息的提取。

图1显示了我们在这里构建的 SV 管道。我们广泛地将自我验证定义为使用对同一对象 的多次调用 LLM 验证其输出，并将每个元件接地它的输出有证据。我们的 SV管道包含四个步骤，每个都以不同的提示调用相同的 LLM。一、*Original*提取步骤直接查询LLM获取所需信息。接下来，省略步骤发现输出中缺少元素，证据步骤依据输出中的每个元素到输入中的文本范围，以及修剪步骤删除输出中不准确的元素。综上所述，我们证明这些步骤可以改善提取信息的可靠性。

此外，SV 为每个输出，以输入中的短文本范围的形式。在可解释性在 NLP 中有多种形式，包括事后特征重要性（Lundberg & Lee，2017；Ribeiro 等人。，2016），本质上可解释的模型（Rudin，2019；辛格等人。，2022a），并可视化模型中间体，例如。注意力（Wiegreffe & Pinter，2019）。可解释的我们产生的基础直接来自法学硕士，类似

[*]同等贡献[1]微软研究院。通讯至：泽拉莱姆·盖罗zelalemgero@microsoft.com>, 钱丹·辛格 <chansingh@microsoft.com >.

[1]所有代码均可在 ‡ github.com/microsoft/clinical-自我验证。

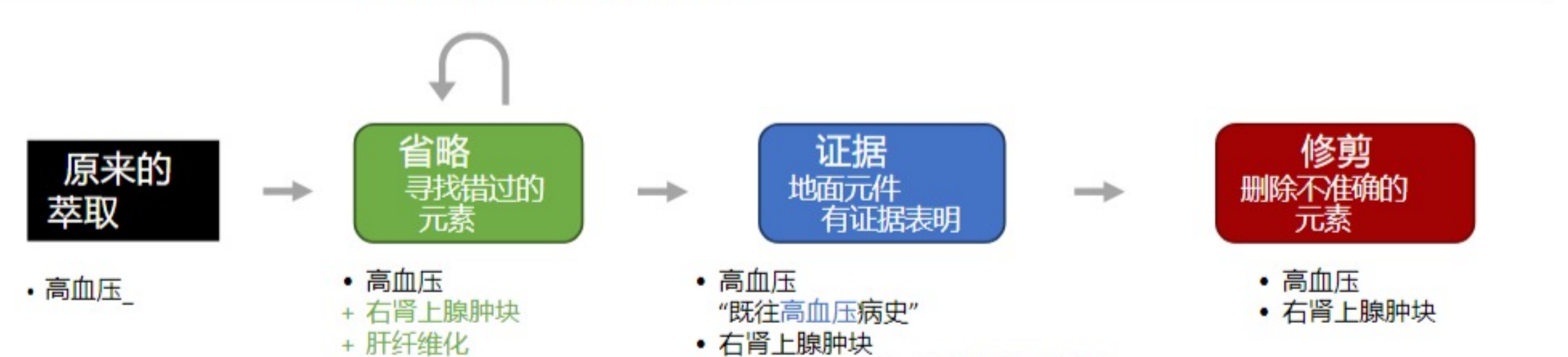

图 1.临床信息提取的自我验证流程概述。每个步骤都会使用不同的提示 调用相同的LLM提炼前面步骤中的信息。在每个步骤下面，我们显示了用于提取指定诊断 列表的缩写输出来自临床记录样本。

最近使用法学硕士来生成解释的作品（Ra-贾尼等人，2019；麦克尼尔等人。，2022；辛格等人。，2023）并将这些解释作为证据（Rashkin 等人，2021；高等人。，2022）。

各种临床信息提取任务的实验以及各种法学硕士，包括GPT-4（OpenAI，2023）和 ChatGPT（Ouyang et al.，2022），显示了 SV 的有效性。在除了提高准确性之外，我们发现提取的解释符合人类对相关信息的判断化，实现人工审计并帮助建立一个可靠地提取临床信息的途径在资源受限的场景中。

2。方法和实验装置

2.1. 方法：自我验证

图1展示了引入的SV的四个不同步骤管道。管道接受原始文本输入，例如。_ _临床的note，并以预先指定的格式输出信息，例如。项目符号列表。它由四个步骤组成，每个步骤都调用相同的法学硕士有不同的提示，以便精炼和将原始输出接地。

原始提取步骤使用特定于任务的提示它指示模型输出可变长度的bul-让列表。在图1的玩具示例中，目标是识别_ _两项诊断为高血压和右肾上腺肿块，但最初的提取步骤仅发现*Hypertension*。

在原始 LLM 提取之后，省略步骤发现输出中缺少元素；在图1 的例子中，它发现右肾上腺肿块和肝纤维化。对于时间较长的 任务输入（平均输入长度大于 2,000 个字符），我们重复遗漏步骤以发现更多潜在的遗漏元素（我们重复五次，并继续重复直到遗漏步骤停止发现新的遗漏）。

接下来，证据步骤将结果 中的每个元素作为基础放入输入中的文本范围。这一步 的接地提供可由人类检查的解释。在图1 的例子中，我们找到了支持第一个的 引用两种诊断，但肝纤维化的引文表明事实上被排除了，因此是一个不正确的诊断姐姐。最后，修剪步骤使用提供的证据来从输出中删除不准确的元素。图1中这个结果消除肝纤维化以返回正确的最终结果列表。总而言之，这些步骤有助于提取准确且可解释的信息。

我们在 Github 中提供了所有步骤中使用的确切提示回购。对于输入较短的任务，我们包括 5 个随机数原始提取提示中的dom数据演示；否则，示例中的所有提示都是固定的。

2.2. 实验装置

数据集表1给出了我们研究的 每项任务的详细信息这里。每个任务都需要提取一个可变长度的 列表元素。在临床试验中手臂提取中，这些是名称不同临床试验组的数据，手动注释 EBM-NLP 数据集（Nye 等人，2018）。在用药方面状态提取任务，除了药物名称之外药物治疗状态必须另外归类为活跃，停产，或者两者都不停产。用于手臂提取 的文本输入 / 药物状态提取相对较小（平均长度分别为 1,620 个字符和 382 个字符）。

对于 MIMIC-III 和 MIMIC-IV（Johnson 等人，2016；2021），我们预测ICD-9或ICD-10代码（对应响应诊断和程序）。我们预测ICD使用所有类型临床记录中相关部分的代码对于MIMIC-III（平均长度：5,200字），但仅显示 MIMIC-IV 的费用摘要（平均长度：1,400字）。ICD代码并不直接出现在文本中- _ _

把，因此任务需要翻译诊断到他们的相关代码。MIMIC数据使用以下方法 进行预处理标准管道（参见附录A.1），我们评估每个任务的 250 个输入的随机子集。

模型我们评估三种不同的模型：GPT-3.5（布朗等人，2020），文本-`davinci-003`，聊天-GPT（Ouyang et al.，2022）gpt -`3.5-turbo`和GPT-4（OpenAI，2023）`gpt4-0314`（聊天模式），全部访问通过 Azure OpenAI API 安全地进行。我们设置了相同的值- LLM 解码的采样温度为 0.1。

评估提取通过不区分 大小写进行评估精确的字符串匹配，我们报告结果宏 F1 分数、召回率和精确率。在某些情况下，这个评估因此评估可能会低估实际表现缩略词或不同名称的存在输出；尽管如此，不同产品的相对表现模型/方法仍应保留。以下com-在平时的实践中，我们将 ICD 代码评估限制在前 50 名数据集中出现的代码。

3。结果

3.1. 自我验证提高预测性能

表2显示了临床提取性能的结果有或没有自我验证。跨不同型号和任务，SV 始终如一地提供性能改进-ment。性能提升有时相当大 _ 大（例如，GPT-4 在 F1 中显示出超过 0.1 的改进用于临床试验手臂拔除和超过 0.3 的改进药物状态提取），以及平均 F1 模型和任务之间的改进为 0.056。我们也与我们连接提示的基线进行比较跨不同的步骤进入一个大提示，即然后用于进行一次LLM调用以获取信息牵引力。我们发现这个 大提示基线执行比表2中报告的基线稍差，使用简单的提取提示（参见比较子详细信息见表A5）。

对于短输入的任务，我们发现 GPT-3.5 的表现最好的，甚至超越 GPT-4，正如在一些例子中所看到的那样最近的作品（例如Patil 等人 2023）。对于MIMIC任务_输入越大，GPT-4表现最好。事实上，GPT-3.5 在 ICD 代码提取方面表现非常差，也许是因为这项任务不仅需要从中提取诊断输入文本，但也知道诊断之间的映射鼻子和 ICD 代码。

表3包含显示不同自我如何验证模块会影响结果。省略步骤 _ 找到缺失的元素，这会增加召回率 ，但降低精度的成本。相比之下，修剪步骤 _

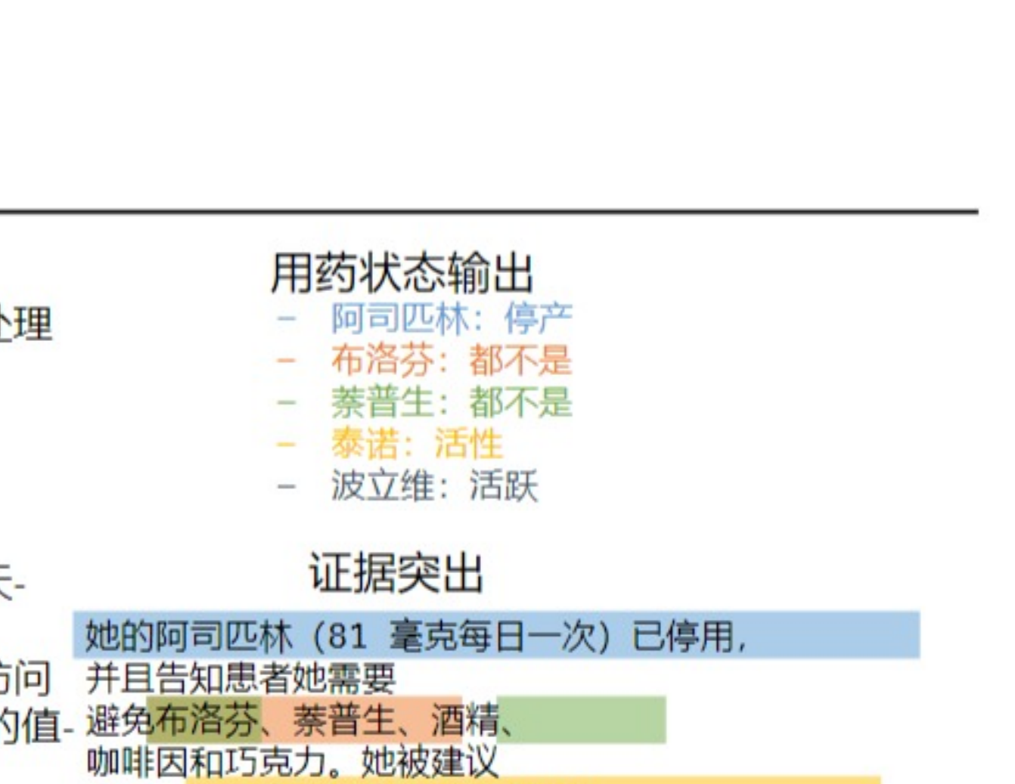

图 2.药物状态的示例输出和解释。对于输出列表的每个元素，我们的管道输出文本 span 包含生成的输出的证据（显示并突出显示）。

（包含证据步骤中的跨度）删除外来元素，从而提高精度。一起（*Full SV*），这些步骤实现了提高F1的平衡。对于具有较长输入的任务（例如MIMIC-IV ICD-10），省略步骤似乎可以提供更多的改进 F1，可能是因为它能够找到被遗漏的证据通过单个提取步骤。

3.2. 自我验证产生解释

图2显示了自验证的示例输出药物状态管道（底层模型是 GPT-4）。在示例中，管道正确识别了每个药物及其相应的状态。此外，_ 管道提供了作为证据的文本范围对于每种退回的药物（以突出显示的方式显示）。这种突出显示使得人类能够进行有效的审计每个元素。在人机循环环境中 ，人类还可以看到元素的结果/亮点修剪，以快速检查是否有任何错误。

表4评估了 SV 提供的证据跨度在先前的工作中收集的人类判断（Nye 等人，2018）。人工审阅者在原文 中注释了跨度与干预措施相对应的文本，其中包括临床临床试验组作为一个子集。表4给出了生成的证据范围与提供的范围重叠由人类注释者。该分数相当大，例如。GPT-4 为 93%。在基线上，人类注释者识别的信息较少超过 3.7% 的代币作为干预措施，因此这些跨度超过圈数精度远高于随机预期机会。

表 1.此处研究的任务和相关数据集。

任务	数据	输出示例
ICD码提取 I(CD-9 和 ICD-10)	250 MIMIC III 报告（Johnson 等人，2016 年）， 250 MIMIC IV 出院摘要（Johnson 等人，2021 年）	[205.0、724.1、96.04]
临床试验手臂拔除	EBM-NLP 摘要的 100 个注释（Nye 等人，2018）	[异丙酚、氟哌利多、安慰剂]
Medication status extraction	105 Annotations (Agrawal et al., 2022) to snippets from CASI (Moon et al., 2012)	{aspirin: discontinued, plavix: active}

Table 2. F1 scores for extraction with and without self-verification (SV). Across different models and tasks, SV consistently provides a performance improvement, although it is sometimes small. Bolding shows SV compared to original, underline shows best model for each task. Averaged over 3 random seeds; error bars show the standard error of the mean.

	ChatGPT	GPT-4	GPT-3.5
Clinical trial arm (Original / SV)	0.342 ±0.010 / **0.456** ±0.007	0.419 ±0.008 / **0.530** ±0.010	0.512 ±0.009 / **0.575** ±0.003
Medication name (Original / SV)	0.892 ±0.004 / **0.898** ±0.002	0.884 ±0.003 / **0.910** ±0.001	0.929 ±0.002 / **0.935** ±0.001
MIMIC-III ICD-9 (Original / SV)	0.593 ±0.003 / **0.619** ±0.005	0.652 ±0.02 / **0.678** ±0.007	0.431 ±0.03 / **0.435** ±0.01
MIMIC-IV ICD-9 (Original / SV)	0.693 ±0.04 / **0.713** ±0.005	0.718 ±0.03 / **0.755** ±0.004	0.691 ±0.02 / **0.702** ±0.02
MIMIC-IV ICD-10 (Original / SV)	0.448 ±0.04 / **0.464** ±0.003	0.487 ±0.02 / **0.533** ±0.002	0.434 ±0.03 / **0.442** ±0.01

Table 3. Ablation results when using different combinations of self-verification steps for two tasks. Omission increases Recall and Prune increases Precision. Together they increase both, improving F1. Evidence improves F1 for Medication Status. Underlying model is the best model for each task (GPT-3.5 for *Medication name* and GPT-4 for *MIMIC-IV ICD-10*). Averaged over 3 random seeds; error bars are standard error of the mean.

	Medication name		
	F1	Precision	Recall
Original	0.929±0.002	0.929±0.003	0.928±0.003
+ Omission	0.913±0.001	0.881±0.003	**0.946**±0.001
+ Prune	0.932±0.002	**0.949**±0.002	0.916±0.003
+ Full SV	**0.935**±0.001	0.942±0.002	0.928±0.001
	MIMIC-IV ICD-10		
	F1	Precicion	Recall
Original	0.487±0.002	0.544±0.003	0.448±0.002
+ Omission	0.517±0.003	0.553±0.003	**0.501**±0.004
+ Prune	0.504±0.004	**0.557**±0.005	0.451±0.003
+ Full SV	**0.533**±0.002	**0.558**±0.002	0.498±0.002

4. Discussion

Self-verification constitutes an important step towards unlocking the potential of LLMs in healthcare settings. As LLMs continue to generally improve in performance, clinical extraction with LLMs + SV seems likely to improve as well.

One limitation of SV is that it incurs a high computational cost as multiple LLM calls are chained together; however, these costs may continue to decrease as models become more efficient (Dao et al., 2022). Another limitation is that LLMs and SV continue to be sensitive to prompts, increasing the need for methods to make LLMs more amenable to prompting (Ouyang et al., 2022; Scheurer et al., 2023) and to make finding strong prompts easier (Shin et al., 2020; Xu et al., 2023; Singh et al., 2022b).

Table 4. Evaluating evidence spans provided by the self-verification pipeline with human-annotated spans. Averaged over 3 random seeds; error bars show standard error of the mean.

	Span overlap accuracy	Span length
GPT-4	0.93 ±0.02	8.20 ±0.48
GPT-3.5	0.84 ±0.03	7.33 ±0.47

Finally, SV can be harnessed in a variety of ways to improve clinical NLP beyond what is studied here, *e.g.* for studying clinical decision rules (Kornblith et al., 2022), clinical decision support systems (Liu et al., 2023), or improving model distillation (Wu et al., 2023; Toma et al., 2023).

References

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chase, H. Langchain: Building applications with llms through composability. `https://github.com/hwchase17/langchain`, 1 2023.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Edin, J., Junge, A., Havtorn, J. D., Borgholt, L., Maistro, M., Ruotsalo, T., and Maaløe, L. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. *arXiv preprint arXiv:2304.10909*, 2023.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., et al. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*, 2022.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. Pal: Program-aided language models, 2023.

Gutiérrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., and Su, Y. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*, 2022.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng IV, S. Mimic-iv-ed. *PhysioNet*, 2021.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jojic, A., Wang, Z., and Jojic, N. Gpt is becoming a turing machine: Here are some ways to program it, 2023.

Kornblith, A. E., Singh, C., Devlin, G., Addo, N., Streck, C. J., Holmes, J. F., Kuppermann, N., Grupp-Phelan, J., Fineman, J., Butte, A. J., et al. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *PLOS Digital Health*, 1(8):e0000076, 2022.

Laursen, M., Pedersen, J., Hansen, R., Savarimuthu, T. R., and Vinholt, P. Danish clinical named entity recognition and relation extraction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 655–666, 2023.

Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., McCoy, A. B., Sittig, D. F., and Wright, A. Assessing the value of chatgpt for clinical decision support optimization. *MedRxiv*, pp. 2023–02, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ma, Y., Cao, Y., Hong, Y., and Sun, A. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*, 2023.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., and Huang, Z. Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*, pp. 37–39, 2022.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Moon, S., Pakhomov, S., and Melton, G. Clinical abbreviation sense inventory. 2012.

Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I. J., Nenkova, A., and Wallace, B. C. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, pp. 197. NIH Public Access, 2018.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Paranjape, B., Lundberg, S., Singh, S., Hajishirzi, H., Zettlemoyer, L., and Ribeiro, M. T. Art: Automatic multi-step reasoning and tool-use for large language models, 2023.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models, 2022.

Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., and Reitter, D. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools, 2023.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Singh, C., Askari, A., Caruana, R., and Gao, J. Augmenting interpretable models with llms during training. *arXiv preprint arXiv:2209.11799*, 2022a.

Singh, C., Morris, J. X., Aneja, J., Rush, A. M., and Gao, J. Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*, 2022b.

Singh, C., Hsu, A. R., Antonello, R., Jain, S., Huth, A. G., Yu, B., and Gao, J. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

Toma, A., Lawler, P. R., Ba, J., Krishnan, R. G., Rubin, B. B., and Wang, B. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.

Wang, B., Deng, X., and Sun, H. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2714–2730, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.174`.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49, 2018.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

Wu, T., Terry, M., and Cai, C. J. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2022.

Xu, B., Wang, Q., Mao, Z., Lyu, Y., She, Q., and Zhang, Y. *k* nn prompting: Beyond-context learning with calibration-free nearest neighbor inference. *arXiv preprint arXiv:2303.13824*, 2023.

Zhang, H., Liu, X., and Zhang, J. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*, 2023.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–375, 2007.

A. Appendix

A.1. Dataset details

MIMIC To preprocess MIMIC data, we follow the steps used by (Edin et al., 2023). For MIMIC IV, we use the available discharge summaries for each patient while we retrieve more relevant sections from other types of clinical notes for MIMIC III. See the code on Github for complete details.

During LLM extraction, we find that directly extracting ICD codes with an LLM is difficult. Instead, we use the LLM to extract diagnoses, and then postprocess them at the end by asking the LLM to convert each diagnosis to its corresponding ICD code.

Clinical trial arm dataset We manually annotate the clinical trial arms from the first 100 abstract in EBM-NLP (Nye et al., 2018) without the use of any LLMs. All annotations are made available on Github. The mean number of extracted clinical trial arms is 2.14, the maximum is 5 and the minimum is 1.

A.2. Extended extraction results

Table A5. F1 scores for two tasks extracted using a single prompt which concatenates all steps in the SV pipeline. Results are slightly worse than the original extraction presented in Table 2. The prompt contains a paragraph similar to the following: *Before you provide your final response:\n(1) Find any medications in the patient note that were missed.\n(2) Find evidence for each medication as a text span in the input.\nn(3) Verify whether each extracted medication is actually a medication and that its status is correct.* Averaged over 3 random seeds; error bars are standard error of the mean.

	ChatGPT	GPT-4	GPT-3.5
Clinical trial arm, original	0.316 ±0.006	0.420 ±0.009	0.436 ±0.008
Medication name, original	0.758 ±0.003	0.850 ±0.016	0.913 ±0.002