

# IAI at #SMM4H-HeaRD 2025: Benchmarking PLMs for medical language understanding tasks

Aman Sinha

Université de Lorraine, Nancy, France  
aman.sinha@univ-lorraine.fr

## Abstract

Recent progress in large language models have demonstrated significant performance improvements across various language tasks. However, their high computational demands pose challenges for deployment in resource-constrained environments. This paper examines the performance of small language models, particularly domain-specific pre-trained language models (PLMs), in the context of healthcare-related language tasks introduced in the SMM4H-HeaRD 2025 shared tasks<sup>1</sup>. We focus on two tasks: detecting dementia family caregivers on Twitter (Task 3) and identifying insomnia in clinical notes (Task 4). Our study primarily utilizes pre-trained language models and investigates the conditions under which these models struggle. The findings offer insights into the limitations of pre-trained models for clinical language understanding, highlighting potential factors that could inform strategies for improving model performance in practical, resource-limited settings.

## Introduction

The growing dominance of large language models (LLMs) is increasingly challenging the traditional use of smaller models in natural language processing. On one hand, LLMs demonstrate remarkable performance across a wide range of language understanding tasks, both in general and domain-specific settings. On the other hand, their substantial computational requirements pose significant obstacles to practical deployment, particularly for institutions with limited resources. While LLMs are currently the preferred choice for achieving high performance in clinical and language understanding tasks, this shift also opens important research avenues for exploring the performance gap between large and small models. Klein et al. (2025) presents a suite of medical language understanding tasks, with a focus on social media and clinical datasets.

In this paper, we present our study on two of the tasks under the SMM4H-HeaRD 2025, namely task 3-detection of dementia family caregiver posts on Twitter, and task 4-detection of insomnia in clinical notes. For our experiments, we utilize various domain specific small language models,

in particular, pretrained language models (PLMs) and study what are the situations where existing PLMs struggle.

## Related Work

Within the healthcare domain, domain-specific adaptations of the Bidirectional Encoder Representations from Transformers (BERT) architecture, such as BioBERT and ClinicalBERT (Lee et al. 2020; Li et al. 2022), were developed to address the inherent complexities of clinical language. The advent of these models specifically targeted the distinctive challenges presented by medical text, which is characterized by intricate medical terminology, lexical ambiguity, and inconsistencies in linguistic usage.

LLMs have shown remarkable proficiency in a wide spectrum of natural language understanding tasks, highlighting their pivotal role in healthcare. However, challenges persist, including the need for robust training data, bias mitigation, data privacy and –most notably– adequate computation infrastructure (Nazi and Peng 2024). These limitations highlight the need to advance smaller, efficient models that offer competitive performance and are feasible for deployment in resource-limited clinical settings.

## Task Description

### Task 3 - Detection of dementia family caregivers on Twitter

The dataset<sup>2</sup> comprised of tweets from users mentioning their experiences with someone in their family suffering from dementia. The task is to identify tweets that report having a family member with dementia (annotated as 1) and not merely mention dementia (annotated as 0). The dataset consists of three splits : train (6,724), validation (353) and unlabeled test (1,769). The training set has a label imbalance where the positive label occurs 67.4% times compared to the negative labels.

### Task 4 - Detection of insomnia in clinical notes

The organizers provided a collection of electronic health records (EHRs) from MIMIC-III Clinical Database (v1.4) (Johnson et al. 2016). The clinical reports are annotated by experts of symptoms of Insomnia. The dataset consisted of 3

Definition 1	Definition 2
<i>Trouble initiating sleep</i>	<i>Fatigue or malaise</i>
<i>Trouble maintaining sleep</i>	<i>Behavioral problems such as hyperactivity, impulsivity, or aggression</i>
<i>Waking up earlier than desired</i>	<i>Impaired attention, concentration, or memory</i>
<i>An explicit mention of Insomnia</i>	<i>Impaired social, family, occupational, or academic performance</i>
	<i>Mood disturbance or irritability</i>
	<i>Daytime sleepiness</i>
<b>Rule A</b>	<i>The patient has insomnia if they meet both <b>Definition 1</b> and <b>Definition 2</b></i>
<b>Rule B</b>	<i>If the patient is prescribed any of the following meds: Estazolam, Eszopiclone, Flurazepam, Lemborexant, Quazepam, Ramelteon, Suvorexant, Temazepam, Triazolam, Zaleplon, Zolpidem</i>
<b>Rule C</b>	<i>If the patient is prescribed any of the following meds: Acamprosate, Alprazolam, Clonazepam, Clonidine, Diazepam, Diphenhydramine, Doxepin, Gabapentin, Hydroxyzine, Lorazepam, Melatonin, Mirtazapine, Olanzapine, Quetiapine, Trazodone. OR any symptoms from <b>Definition 1</b> or <b>Definition 2</b></i>

Table 1: Description of the Insomnia rules for Task 4 (Subtask 2A)

splits: 70 train samples, 20 validation samples, and 100 test samples. The shared task 4<sup>3</sup> consists of two subtasks:

**Subtask1** involves a binary classification for identifying the clinical note for the patient suffering with the condition of Insomnia. The training set, here, suffers from an imbalance between positive(“yes”) and negative(“no”) classes with as 61.4% of the samples were positive.

**Subtask2** is further divided into two parts, namely Subtask 2A, which is a multilabel classification task which involves evaluating each clinical note against the defined Insomnia rules (see table 1): Definition 1, Definition 2, Rule A, Rule B, and Rule C. Subtask 2B extends Subtask 2A by requiring the identification and extraction of text evidence from the clinical note that supports each classification.

For Task 4, we participated and report only for Subtask 1 and Subtask 2A tasks.

## Experimental Setup

### Models

We considered existing pre-trained language models (PLMs) for our study for both the shared tasks.

**Task 3:** Following (Klein et al. 2022), we considered BERTweet (Nguyen, Vu, and Nguyen 2020), TwHIN-BERT (Zhang et al. 2022) and BERT as our baseline.

**Task 4:** We considered various biomedical specialized pre-trained language models including BioBERT (Lee et al. 2020), PubmedBERT (Gu et al. 2021), SciBERT (Beltagy, Lo, and Cohan 2019), MedBERT (Vasantharajan et al.

2022), ClinicalBigBird (Li et al. 2022) and BioClinicalBERT (Alsentzer et al. 2019) for our experiments. For subtask 2A, for Definition 1 and 2, we used language models, however, we used logical operator module for Rule A, a fuzzy string matching module for Rule B and C.

### Loss function

In order to address the imbalance nature of dataset in both the tasks we participated in, we used focal loss (Lin et al. 2017) to train all of our models. Focal Loss enables the standard cross-entropy loss designed to address class imbalance by focusing on “hard” examples, while reducing the loss for “easy” examples.

$$L_{\text{focal}} = -\alpha_t(1 - \hat{p}_{i,y_i})^\gamma \log(\hat{p}_{i,y_i})$$

where  $\gamma$  is the focusing parameter to reduce the contribution of “easy” examples (default set to 2.0), and  $\hat{p}_{i,y_i}$  is the softmax probability for the correct class  $y_i$ .

### Evaluation

**Task 3:** For evaluating the binary classification we used F1 score with a average setting of binary.

**Task 4:** For evaluating subtask 1, which was a binary classification task, we used micro-average F1 score. For subtask 2A, which was a multi label classification task we again used micro-average F1 score as the evaluation strategy.

### Reproducibility

All the models are trained with the common setting on 20 epochs with early stopping, with a patience of 3 based on either evaluation metric F1 score or loss function criterion. The learning rate is set to  $1e-5$ . And, the choice of optimizer was AdamW (Loshchilov and Hutter 2017).

<sup>3</sup><https://github.com/guilopgar/SMM4H-HeaRD-2025-Task-4-Insomnia>

model	Val-F1
BERT	0.925
BERTweet	0.946
TwHIN-BERT	0.929

Table 2: Results on validation set for task 3

	F1	P	R
MEAN	0.885	0.925	0.892
MEDIAN	0.953	0.946	0.969
<i>Our submissions</i>			
BERTweet	<b>0.964</b>	<b>0.956</b>	0.971
Ensemble	0.953	0.934	<b>0.974</b>

Table 3: Final results on test set for task 3. **Bold** notation denotes highest scores among our submission.

In regard of the dataset size of the two tasks, for Task 3, we trained the models only with one seed. However, for Task 4, we decided to 5-fold cross validation over the training set to tackle the small training split size. Also, it is to be noted that be used only the first 512 token of the clinical notes for training the models. Lastly, for subtask 2A (Definition 1) input format was as follows:

Trouble initiating sleep OR Trouble maintaining sleep OR Waking up earlier than desired OR An explicit mention of Inso [SEP] <<clinical note>>.

### Our Final Submission

For the final submissions over the test set for both the tasks, we submitted ensemble of multiple models, by aggregating prediction via majority voting strategy.

**Task 3:** We submitted two submission based on the scores on validation set (See table 2). We selected one model which was the best among our pool of trained models (BERTweet). For the next one, we performed ensembling over all the three models’ prediction and submitted an ensemble solution.

**Task 4:** We submitted three submissions based on top scores on validation set (See table 4) for subtask 1 and subtask 2A. Firstly, for both, we selected the top 10 models from the cross-validation pool of models. And, then for each of the subtasks, we submitted ensembles the predictions of top-3, 5, and all 10 models.

## Results

**Task 3:** We show our experiment results with the validation set in table 2 and the final leaderboard statistics in table 3. We observe that BERTweet outperforms BERT and TwHIN-BERT models by ~2%. We also notice that TwHIN-BERT obtains slight better performance compared to the vanilla BERT models. On test set, our final submitted systems outperforms the mean and median submission.

model	Subtask 1			Subtask 2a		
	F1	P	R	F1	P	R
Clinical BigBird	0.804	0.704	0.950	0.664	0.824	0.557
BioBERT	0.758	0.680	0.867	0.666	0.771	0.590
SciBERT	0.796	0.706	0.917	0.674	0.722	0.648
BioClinicalBERT	0.764	0.687	0.867	0.663	0.747	0.609
PubMedBERT	0.765	0.699	0.850	0.636	0.716	0.581
MedBERT	0.815	0.728	0.933	0.625	0.639	0.623

Table 4: Results on validation set for task 4 (average result of 5-fold cross validation)

	Subtask 1			Subtask 2a		
	F1	P	R	F1	P	R
MEAN	0.877	0.853	0.913	0.717	0.673	0.788
MEDIAN	0.869	0.840	0.935	0.692	0.650	0.818
<i>Our submissions</i>						
Ens. (Top 3)	0.868	0.811	<b>0.935</b>	0.681	0.646	0.719
Ens. (Top 5)	0.868	0.811	<b>0.935</b>	<b>0.689</b>	<b>0.681</b>	0.697
Ens. (Top 10)	<b>0.875</b>	<b>0.840</b>	0.913	0.681	0.640	<b>0.727</b>

Table 5: Final results on test set for task 4. **Bold** notation denotes highest score among our submission.

**Task 4:** We show our experiment results on validation set in table 4 and the final leaderboard statistics in table 5. For Subtask 1, we observe that MedBERT, Clinical BigBird, and SciBERT have comparable performance and they outperform the three models by a good margin of ~3-6%. For Subtask 2A, we notice that SciBERT is able to outperform the other models however the margin is not very significant. On the test, we notice similar performance compared to validation set. For subtask 1, our submission of Ensemble of top-10 models from validation set outperforms median scores and is comparable to mean scores. However, for subtask 2A, our best solution Ensemble (Top5) is only comparable with median solution.

## Error Analysis & Discussion

**Task 3:** Firstly, we observe that both BERTweet and TwHIN-BERT, which are pre-trained on tweets, outperform BERT, which is trained on general-domain data. The key difference between BERTweet and TwHIN-BERT lies in their pretraining corpora: BERTweet was trained on 5 million English COVID-related tweets, while TwHIN-BERT was trained on multilingual data spanning over 100 languages. This gives BERTweet an advantage over TwHIN-BERT, as its pretraining data is more closely aligned with the task. To find more insights, we select all the 15 tweets for which each of the three models (ref table 2) predicted wrong class and then, we compare the gold label with human and LLM (GPT-4o) judgment. table 6) shows the samples where there is disagreement between gold labels and LLMs (GPT-4o) and corresponding human judgment. We notice that, GPT as well struggles for 46% cases out of the 15 tweets. This can be attributed to its incapability to differentiate between *intonation of the utterance* (See in Table 6, §6, 12) often interpreting the entire text as factual – this behavior can be

	Tweets	GOLD	Human	GPT
3	While I do appreciate the alternatives to guardianship, “g” can be very important for some. When <b>UNcle John started fighting dementia</b> . He would become combative refusing service. Dr’s wouldn’t talk with my mom forcing her to gain guardianship. Expensive & ridiculous process! <a href="https://t.co/JET0c1TL7o">https://t.co/JET0c1TL7o</a>	0	1	1
6	@FruiKace This has happened to my mom a many times. Whenever <b>my dad says my mom</b> is forgetting stuff and <b>has dementia</b> I ask if she has a UTI.	0	0	1
8	@aja1979 @AnnelieseDodds @UKLabour Yvette Cooper was Min of Work & Pensions when <b>my husband</b> was dying. She forced sick ppl to go to interviews & give reasons for not working. He was <b>suffering from vascular dementia</b> amongst other things & thought he had done wrong; his confusion & tears were heart-breaking	0	1	1
9	I’m currently off work due to poor MH. Too many stresses at the same time, has led to burnout- Dad in end stage of life, <b>Mum with undiagnosed dementia</b> and self-neglect, other family stuff in Ireland I cant discuss and daughter’s health. I have no idea when I will return to work☺	0	1	1
10	In my, ‘Meetings,’ with Dementia, if it helps, ergo-do it! Again, one size does not fit all. For ex, <b>Mum</b> , was content just <b>watching TV</b> . Think, brought her in the present, no need to remember the past?	0	0	1
12	@LokeyOrbe @JudyLaTorre6 @mtgreenee Darling, after you’ve been married for that many years, he deserves a pass!! Lol. You know it’s not easy putting up with a strong woman!! Dementia is not easy. I’ve <b>attempted to diagnose</b> my mother with it, but the old troll still knows where she hides every piece of silver.♥	0	0	1
13	Joe Biden literally starts blanking on national TV in one of the most progressive cases of dementia I’ve ever seen ( <b>my grandpa had it</b> and it got bad FAST) and the mainstream media is covering for him. Trump is right they are the enemy of the people.	0	0	1

Table 6: Comparison between GOLD labels and Human and LLM for failed cases in validation set for Task 3. **Highlighted** words denote the reasoning for LLM’s decision for detecting dementia caregiver tweets.

extended PLMs as well. Additionally, language models have difficulty accounting for the temporality factor, i.e., whether the statement refers to the past, as seen in samples §3, 8, and 13. Finally, there are cases such as sample §9 that hinge on the certainty factor of the event. When the task definition is ambiguous, this can lead to human disagreement during annotation, making the task inherently challenging for any model, regardless of its benchmark performance.

**Task 4:** Firstly, for subtask 1, we observe that Clinical Bigbird and MedBERT outperform other models. This can be attributed to their pretraining datasets: Clinical BigBird is trained on MIMIC-III, while MedBERT uses n2c2, BioNLP, and CRAFT community datasets. Further, during error analysis, we identified two samples in the validation set for which all the PLM models (See table 4, Subtask 1) consistently failed to predict the correct class. We additionally prompted GPT on these two cases and obtained 100% agreement with the gold labels. Phrases such as “*very tired and weak*”, “*nightmares*” in the first example and “*neuro / short deficit*”, “*very cooperative, sweet, now at times angry*” and “*discomfort and incisional pain*” in second sample (See table 7) can loosely be associated with Insomnia. This suggests that the models maybe marking even any non-specific symptoms even if it’s not strictly constrained to explicit mentions or direct evidence as an indication of Insomnia. As shown in table 4 and 5, the recall (R) of all the PLMs is  $\geq 85\%$ , implying the presence of false-positives. This indicates that the PLMs are more cautious and tend to prioritize

identifying most or all actual cases of insomnia, even at the cost of precision.

For subtask 2A, the subtask is composed of multiple sub-components (Definition 1, Definition 2, Rule A, Rule B and Rule C) which have a non-zero overlap (See table 1) but are evaluated collectively. As a result, an incorrect prediction in any of the sub-components can lead to an overall drop in performance. This is evident in the lower overall performance compared to Subtask 1, which can also be attributed to the complexity of Subtask 2A as it involves identifying symptoms in the clinical notes which not necessarily are mentioned explicitly. Furthermore, since Rule A and C are dependent on Definition 1 and 2, they are directly effected by models’ performance on each of the respectively.

During error analysis, we observed that for Definition 1, the mean accuracy was 67% with 4 out of the 20 cases where all the PLMs failed to predict the correct class. For Definition 2, the mean accuracy dropped to 58% with 6 out of the 20 cases where all models failed. Additionally, only 23% of the training samples for Rule A were positive, which further impacted model generalization.

On the other hand, for Rule B, we obtained 100% accuracy over validation set as it involved only fuzzy string search for the primary medicines mentioned in the Rule B. However, fuzzy matching alone was insufficient for Rule C, as it failed to capture all relevant cases. We also found out that in  $\geq 25\%$  of the failure cases, the clinical note length exceeded 1000 words—potentially causing truncation and

Clinical Report		Insomnia	GPT
Subtask 1			
10	female patient in eighties prescribed Digoxin, Hydromorphone, Propranolol LA, Sodium Chloride 0.9% Flush, Heparin, Levofloxacin, Vancomycin HCl, Iso-Osmotic Dextrose, Dextrose 5%, Furosemide, NS, Metronidazole, Potassium Chloride, Acetaminophen, Metoprolol, Magnesium Sulfate, Miconazole Powder 2% mental status: alert oriented. obeys commands.pt slept most of night. cv: bp up to 175-180 with discomfort.bp decrease to 120-130 when comfortable.,hr perm pacer av pacing at 72. gu: urine output low. 20-25 cc/hr. urine output dropped off to 10 cc about 0200. treated with 500 cc ns urine picked up to 20-25 cc/hr gi: pos bowel sounds abd soft. small amount of soft stool. integumentary: buttocks and r breast very pink rash. needs mycostatin .spoke with team and asked for order for mycostatin powder. right breast is macerated..skin is peeling. resp: pt has clear upper airways ,diminished at bases. coughing and raising small amounts white.o2 at 2 liters nc, chest tube draining sm amounts serosanguinous. pt is <b>very tired and weak</b> . she is uncomfortable but does not want any med stronger than tylenol.because she says it gives her <b>nightmares</b> . tylenol times 2 with fair effect.	no	no
12	female patient in sixties prescribed no drugs 7a-7p NPN s: I hurt everywhere o: see carevue for all objective data neuro: pt c/o no visual disturbances, MAE, strength =, pt w/ some <b>short term deficits</b> . Family states pt at baseline is <b>very cooperative, sweet, now at times angry</b> , but is cooperative. Neuro here for consult. Head CT done. cv: hemodynamically stable w/ hr 70-80's sr, pacing wires attached. bp 130-160/50-70. resp: sats 92-95 on RA, lungs w/ crackles at bases. id: tm 99.0 po heme: hct 23.4 this am, HO aware. end: bs 130-234, covered per riss. gu: foley draining cl yel urine 100cc/hr gi: taking sm amts soft food, no stool today. skin: chest and leg incisions d/i pain: c/o general <b>discomfort and incisional pain</b> , given tylenol and percocet w/ relief. activity: oob to chair w/ 2 assista, walked 30' w/ PT using wheeling walker, around rm w/ rn. social: 3 daughters in to visit, met w/ social worker <b>[[Name (NI)]]</b> <b>[[Last Name (NamePattern1)]]</b> , case manager <b>[[Name (NI) 346]]</b> <b>[[Last Name (NamePattern1)]]</b> , Dr <b>[[Last Name (STitle) 10415]]</b> in to speak w/ family and pt. A: neuro changes s/p CABG P: Monitor neuro status, goal keep sbp >130. Continue cardiac rehab.	no	no

Table 7: Examples from Task 4 (Subtask 1) validation set where every PLM failed to predict the correct class. **Highlighted** words denote the reasoning for PLM’s decision for mislabelling the clinical report for Insomnia by GPT.

loss of key information. Finally, an important contributing factor to the overall lower performance is the small number of training samples (70), particularly for Definition 2 which also suffered with class imbalance.

### Conclusion

In this work, we present the details of our participation in Task 3 and Task 4 (subtask 1 and 2A) from SMM4H shared task challenge. With our participation, we aim to identify the limitation of pretrained language models for healthcare-related language understanding tasks on social and clinical datasets. Overall, we observe domain specific pre-trained data is likely to help models when fine tuned for specific tasks with adequate amount of data. Data augmentation techniques; contextual finetuning for better accounting temporality and certainty of medical events can be further explored for the clinical tasks to deal with lack of generalization of models in case of low resource clinical setting.

### References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: Pre-trained Language Model for Scientific Text. In *EMNLP*.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Gryboski, L.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Mazzotti, D.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Klein, A. Z.; Magge, A.; O’Connor, K.; and Gonzalez-Hernandez, G. 2022. Automatically identifying twitter users for interventions to support dementia family caregivers: annotated data set and benchmark classification models. *JMIR aging*, 5(3): e39547.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

- Li, Y.; Wehbe, R. M.; Ahmad, F. S.; Wang, H.; and Luo, Y. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Nazi, Z. A.; and Peng, W. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, 57. MDPI.
- Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.
- Vasantharajan, C.; Tun, K. Z.; Thi-Nga, H.; Jain, S.; Rong, T.; and Siong, C. E. 2022. MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1482–1488.
- Zhang, X.; Malkov, Y.; Florez, O.; Park, S.; McWilliams, B.; Han, J.; and El-Kishky, A. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv preprint arXiv:2209.07562*.