Exploring the Limitations of PLMs in medical language understanding tasks

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

📢 **Looking for Postdoc positions in Healthcare/NLP!**

Introduction

- LLMs excel at many language tasks, but **demand substantial compute to deploy**.
- We study PLMs for medical language tasks and **investigates the conditions under which they struggle**.
- Surprisingly, PLMs are not entirely bad. They just require more contextual fine-tuning.

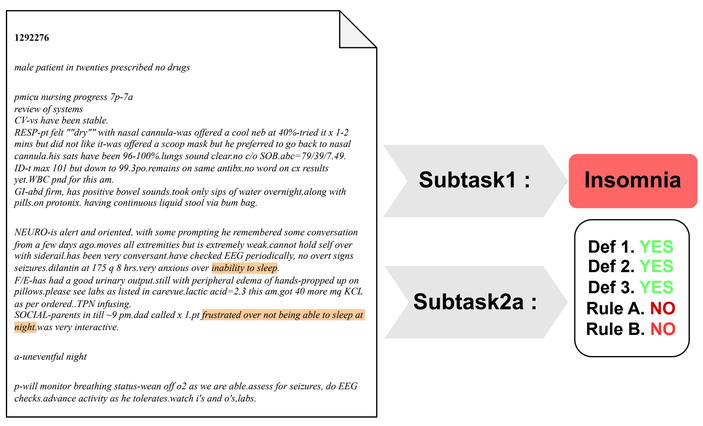
Task Description

- **T3: Dementia caregiver detection in tweets.**

Evelyn has dementia, and I know this. But when she asked me today how my dad was doing ... it hurts. — NO

@kurteichenwald Actually after the hell I went through with my sibs trying to rob my mom with dementia..No. I'm not stunned at all — YES

- **T4: Insomnia detection in clinical notes.**



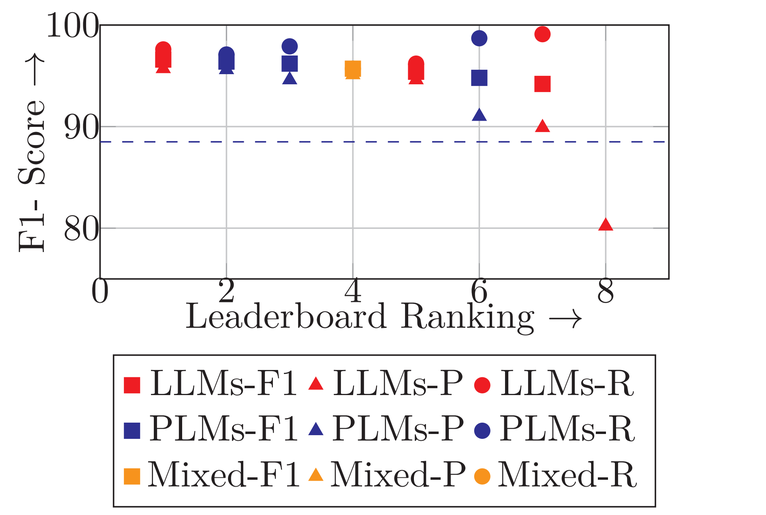
Def 1.	*Trouble initiating sleep* OR *Trouble maintaining sleep* OR *Waking up earlier than desired* OR *An explicit mention of insomnia.*
Def 2.	*Fatigue or malaise.* OR *Impaired attention, concentration, or memory.* OR *Impaired social, family, occupational, or academic performance.* OR *Mood disturbance or irritability.* OR *Daytime sleepiness.* OR *Behavioral problems such as hyperactivity, impulsivity, or aggression.* OR *Decreased motivation, energy, or initiative. Proneness to errors or accidents.* OR *Concerns or dissatisfaction with sleep.*
Rule A.	**Def 1.** AND **Def 2.**
Rule B.	Estazolam, Eszopiclone, Flurazepam, Lemborexant, Quazepam, Ramelteon, Suvorexant, Temazepam, Triazolam, Zaleplon, Zolpidem
Rule C.	Acamprosate, Alprazolam, Clonazepam, Clonidine, Diazepam, Diphenhydramine, Doxepin, Gabapentin, Hydroxyzine, Lorazepam, Melatonin, Mirtazapine, Olanzapine, Quetiapine, Trazodone. OR **Def 1.** OR **Def 2.**

Results: T3: Dementia caregiver detection in tweets.

- **Models.** BERT, BERTweet, TwHIN-BERT
- **Loss Function.** Focal Loss.

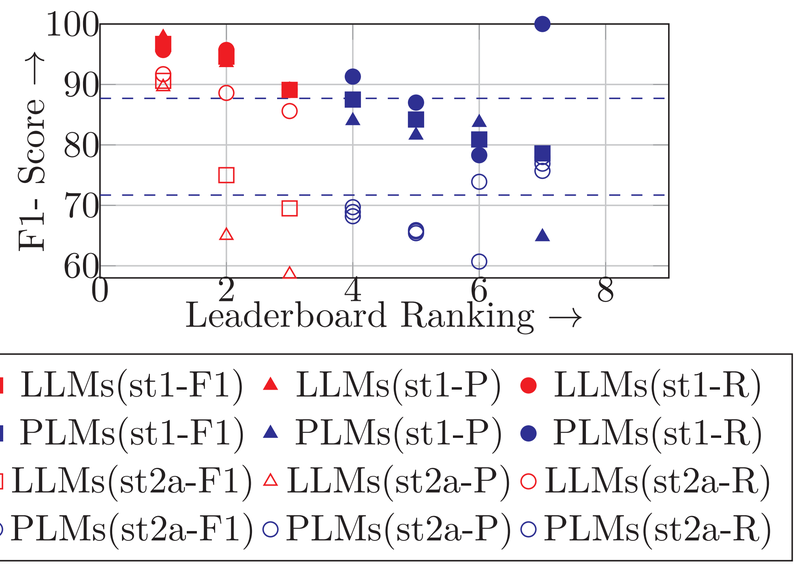
$$L_{\text{focal}} = -\alpha_t (1 - \hat{p}_{i,y_i})^{\gamma} \log(\hat{p}_{i,y_i})$$

- **Best Val Model.** BERTweet
- **Final Model.** BERTweet
- **Overall Rank.** 2nd out of 7 teams.



Results: T4: Insomnia detection in clinical notes.

- **Models.** BioBERT, PubmedBERT, SciBERT, MedBERT, ClinicalBigBird and BioClinicalBERT
- **Loss Function.** Focal Loss.
- **Best Val Model.** MedBERT (st1), SciBERT (st2a)
- **Submitted Model.** Ensemble of top-10 best seeds on validation set.
- **Overall Rank.** 4th out of 7 teams.



Error Analysis: PLMs vs GPT-4o

T3: Dementia caregiver detection in tweet.

- All PLMs failed on 15/353 val. cases; GPT-4o also failed on 7/15.
- Error Attribution:

Utterance intonation	@FruitKace This has happened to my mom a many times. Whenever my dad says my mom is forgetting stuff and has dementia I ask if she has a UTI.
Temporality	Joe Biden literally starts blanking on national TV in one of the most progressive cases of dementia I've ever seen (my grandpa had it and it got bad FAST) and the mainstream media is covering for him. Trump is right they are the enemy of the people.
Certainty	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, Mum with undiagnosed dementia 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

T4: Insomnia detection in clinical notes.

- Low resource dataset

Train: 70, Val: 20 , Test: 100

Subtask 1

- 2 out of 20 val. cases for which all PLMs failed; GPT-4o succeeded.
- Error Attribution:

Extrapolating in-direct symptoms	"very tired and weak", "nightmares", "neuro / short deficit", "very cooperative, sweet, now at times angry" , "discomfort and incisional pain"

Subtask 2a

- Subtask 2a was complex. Interdependent sub-subtasks!
- PLMs struggle +10% on Def 2. compared to Def 1.
- For Rule B. we scored 100% by fuzzy string matching.
- Rule A and C model struggled due to inter-subsubtask dependency.

Takeaways

- PLMs often offer better overall **recall**.
- Domain specificity helps only when models are fine-tuned on specific tasks with enough data.
- Contextual finetuning medical models can improve handling of event time and certainty.

Want to know me or read more?