



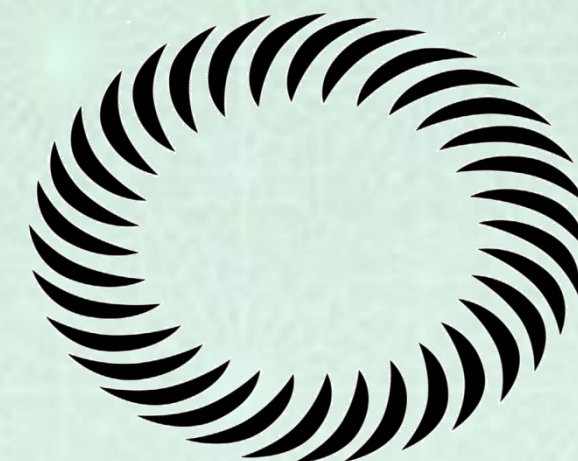
نمایشگاه هفته پژوهش و فناوری  
دانشگاه صنعتی همدان  
آذر 1404

# تنظیم دقیق یک مدل زبانی سبک برای استخراج اطلاعات دارویی

زینب شمال زاده ، دکتر حسن بشیری

گروه مهندسی کامپیوتر، دانشگاه صنعتی همدان

(Bashiri@hut.ac.ir), (zeinabshomalzadeh@gmail.com)



هفته ملی پژوهش و فناوری

## چکیده

استخراج اطلاعات دارویی به دلیل کاربرد گسترده در حوزه‌های زیست‌پزشکی و داروسازی به یک موضوع تحقیقاتی مهم تبدیل شده است. متون زیست‌پزشکی شامل اصطلاحات فنی پیچیده، نمادها و معادلات شیمیایی هستند و به دلیل انعکاس دقیق دانش دارویی می‌توانند تغییرات و روابط بین داروها، بیماری‌ها و تعاملات را به طور عینی نشان دهند. تاکنون مدل‌های متنوعی برای استخراج اطلاعات معرفی شده‌اند که شامل روش‌های مبتنی بر قوانین، یادگیری ماشین، یادگیری عمیق و مدل‌های زبانی بزرگ است. با این حال، محدودیت داده‌های فارسی و چندزبانه، ابهام معنایی اصطلاحات و نیاز به دقت بالا در اطلاعات حساس باعث می‌شود مدل‌ها نتوانند عملکرد مطلوبی در محیط‌های واقعی ارائه دهند. برای حل این مشکل یک مدل زبانی سبک با تنظیم دقیق برای استخراج اطلاعات دارویی مانند شناسایی موجودیت‌های نامدار و روابط پیشنهاد می‌گردد. این مدل با در نظر گرفتن محدودیت‌های محاسباتی و تمرکز بر داده‌های کم‌منبع، تاثیر عوامل زبانی و معنایی را در بهبود استخراج اطلاعات لحاظ می‌نماید. برای پیاده‌سازی روش، ابتدا یک مدل زبانی سبک پیاده‌سازی خواهند شد. سپس رویکرد تنظیم دقیق بر اساس وظایف دارویی ارائه خواهد گردید. انتظار می‌رود مدل پیشنهادی ضمن حفظ دقت قابل قبول، کارایی محاسباتی و سرعت پردازش را در محیط‌های کم‌منبع به‌طور چشمگیری بهبود بخشد و قابلیت استفاده در سامانه‌های واقعی را فراهم آورد.

**واژه های کلیدی:** مدل زبانی سبک، تنظیم دقیق، استخراج اطلاعات دارویی، پردازش زبان طبیعی زیست‌پزشکی

## مقدمه

حجم عظیم متون پزشکی و دارویی منتشرشده، ضرورت استخراج خودکار اطلاعات ساختاریافته از این منابع غیرساختاریافته را بیش از پیش نمایان ساخته است. [۱] اطلاعات دارویی شامل سه مؤلفه اصلی یعنی شناسایی موجودیت‌های نامدار، استخراج روابط دارویی و استخراج رویدادها است. [۲] با این حال، در این پژوهش تمرکز بر دو مؤلفه نخست یعنی شناسایی موجودیت‌های دارویی و روابط میان آن‌ها قرار دارد و به استخراج رویداد پرداخته نمی‌شود، زیرا هدف تحقیق، طراحی مدلی بهینه برای این دو وظیفه کلیدی است. اگرچه مدل‌های زبانی بزرگ مانند BERT و GPT در این حوزه عملکرد چشمگیری داشته‌اند، اما نیاز شدید آن‌ها به منابع محاسباتی، حافظه بالا و زیرساخت‌های پیچیده، کاربرد عملی‌شان را در محیط‌های واقعی محدود می‌سازد. در مقابل، مدل‌های زبانی سبک مانند DistilBERT، TinyBERT و MobileBERT با بهره‌گیری از تکنیک‌هایی چون تقطیر دانش و یادگیری مداوم طراحی شده‌اند تا ضمن کاهش پیچیدگی محاسباتی، دقت مطلوب خود را حفظ کنند. با این حال، اغلب این مدل‌ها بر متون عمومی آموزش دیده‌اند و از دانش تخصصی در حوزه زیست‌پزشکی بی‌بهره‌اند. از این‌رو، در این پژوهش تمرکز بر طراحی و تنظیم دقیق یک مدل زبانی سبک مختص حوزه زیست‌پزشکی است تا بتوان از آن در استخراج دقیق، سریع و کارآمد اطلاعات دارویی در محیط‌های کم‌منبع بهره برد. (شکل ۱)

## اهداف و روش پژوهش

### اهداف اساسی پژوهش:

- تنظیم دقیق یک مدل زبانی سبک با کارایی بالا برای وظایف استخراج اطلاعات دارویی
- به‌کارگیری روش‌های تنظیم دقیق کارآمد متناسب با داده‌های دارویی
- ارزایی عملکرد مدل پیشنهادی در وظایف استخراج رابطه و شناسایی موجودیت‌ها
- مقایسه عملکرد مدل زبانی سبک تنظیم‌شده با نتایج گزارش‌شده مدل‌های زبانی بزرگ از نظر دقت، کارایی محاسباتی و قابلیت کاربرد در محیط‌های عملیاتی کم‌منبع.

### روش پژوهش:

در این پژوهش یکی از مجموعه داده‌های معتبر زیست‌پزشکی مانند BC5CDR یا BioRED انتخاب شده و داده‌ها برای دو وظیفه شناسایی موجودیت‌ها و استخراج روابط پیش‌پردازش می‌گردند. سپس یک مدل زبانی سبک با استفاده از روش‌های تنظیم دقیق بر داده‌های دارویی آموزش داده می‌شود. عملکرد مدل با مجموعه‌ای از معیارهای کمی و تحلیل‌های کیفی مانند بررسی خطاها سنجیده خواهد شد. هدف، توسعه مدلی کارآمد و قابل استقرار در محیط‌هایی است که در آن‌ها توازن میان دقت، سرعت و مصرف منابع اهمیت دارد. (شکل ۲)



شکل ۲: فلوچارت روش تحقیق

## یافته‌های پژوهش

متون زیست‌پزشکی دارای ساختار پیچیده و وابستگی‌های معنایی ظریفی هستند که استخراج اطلاعات از آن‌ها را دشوار می‌کند. مدل‌های زبانی بزرگ با وجود دقت بالا، به دلیل هزینه زیاد اجرایی برای بسیاری از سازمان‌ها قابل استفاده نیستند. به همین دلیل نسخه‌های سبک‌شده مانند DistilBERT و TinyBERT توسعه یافته‌اند تا بخشی از توان مدل‌های بزرگ را با منابع کمتر ارائه دهند.

پژوهش‌هایی مانند مطالعه امید روحانیان نشان داده‌اند که مدل‌های فشرده‌شده—حتی با تعداد پارامتر بسیار کمتر—می‌توانند در وظایف RE و NER عملکردی نزدیک به مدل‌های بزرگ داشته باشند. [۳] این ویژگی برای حوزه دارویی اهمیت زیادی دارد، زیرا مدل باید موجودیت‌هایی مانند دارو، عوارض و بیماری‌ها را شناسایی کرده و روابط میان آن‌ها را استخراج کند. مدل‌های سبک به دلیل مصرف حافظه کمتر و سرعت بالاتر، گزینه‌ای مناسب برای کاربردهای بالینی و سامانه‌های واقعی کم‌منبع به‌شمار می‌روند. [۴]

## راهنمای پیشنهادی

- راهنمای پیشنهادی برای بهبود مدل‌های زبانی سبک در استخراج اطلاعات دارویی با وجود پیشرفت‌های اخیر، بسیاری از مدل‌های سبک همچنان با چالش‌هایی مانند کمبود داده، افت دقت در سناریوهای تخصصی و محدودیت‌های محاسباتی روبه‌رو هستند. بر این اساس، مهم‌ترین راهنمای قابل‌استفاده برای بهبود عبارت‌اند از:
- تنظیم دقیق پارامتر-کارا: بهره‌گیری از روش‌هایی مانند LoRA، Prefix-Tuning و Prompt Tuning که امکان تزریق دانش تخصصی را بدون افزایش محسوس هزینه محاسباتی فراهم می‌کنند.
  - تقویت داده در شرایط کم‌داده: استفاده از روش‌های تولید نمونه‌های مصنوعی یا بازنویسی داده‌ها برای افزایش تنوع و پایداری مدل در برابر نویزهای متنی.
  - ارزیابی چندبعدی: علاوه بر معیارهای کلاسیک، توجه به معیارهایی مانند زمان استنتاج، مصرف حافظه، مقاومت مدل در برابر داده‌های نویزی و توانایی یادگیری از داده‌های کم‌برچسب.
  - تحلیل خطاهای حساس بالینی: تمرکز بر خطاهای پرریسک مانند اشتباه در استخراج عوارض شدید دارویی و طراحی معیارهای ارزیابی حساس به خطر یا استفاده از LLMها به عنوان داور.
- این راهنما می‌تواند دقت، پایداری و قابلیت اعتماد مدل‌های زبانی سبک را در کاربردهای سلامت دیجیتال به شکل معناداری ارتقا دهند.

## منابع

- [1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J., BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Oxford University Press, 2020.
- [2] Yang, J., Walker, K. C., Bekar-Cesaretli, A. A., Hao, B., Bhadelia, N., Joseph-McCarthy, D., & Paschalidis, I. C., Automating biomedical literature review for rapid drug discovery: Leveraging GPT-4 to expedite pandemic response, Elsevier, 2024.
- [3] Rohanian, O., Nouriborji, M., Kouchaki, S., & Clifton, D. A., On the effectiveness of compact biomedical transformers, Oxford University Press, 2023.
- [4] Rohanian, O., Nouriborji, M., Jauncey, H., Kouchaki, S., Nooralahzadeh, F., Clifton, L., & C. Group, I. C., Lightweight transformers for clinical natural language processing, Cambridge University Press, 2024.

## تایید استاد راهنما

نام و امضا استاد راهنما : دکتر حسن بشیری

تایید تحصیلات تکمیلی :

تایید امور پژوهشی: