1) مسئلهٔ اصلی چیست و چرا مهم است؟

این مقاله روی تشخیص شایعات/اطلاعات نادرست حوزه سلامت در فضای وب (به‌طور خاص در زبان چینی) تمرکز دارد. نویسندگان می‌گویند با رشد سرعت انتشار محتواهای حوزه سلامت در اینترنت، «شایعات سلامت» در کنار اطلاعات درست پخش می‌شوند و می‌توانند تهدیدی برای سلامت عمومی باشند. مشکل بزرگ‌تر این است که (۱) اکثر دیتاست‌های موجود برای شایعات سلامت کوچک هستند و تنوع کافی برای آموزش مدل‌های عمیق را ندارند. و (۲) روش‌های سنتی تشخیص شایعه معمولاً فقط برچسب (شایعه/غیرشایعه) را خروجی می‌دهند، بدون اینکه دلیل یا استناد علمی ارائه کنند. برای کاربران نهایی، صرفاً دانستن اینکه یک خبر شایعه است کافی نیست؛ آن‌ها نیاز به دلایل و شواهد علمی دارند تا قانع شوند.

مقاله برای پرکردن این شکاف دو کار اصلی انجام می‌دهد:

• ساخت یک دیتاست بسیار بزرگ به نام HealthRCN با ۱.۱۲ میلیون نمونه داده شایعه/اطلاعات سلامت.

• ارائهٔ یک سیستم به نام HRDE که یک LLM تقویت‌شده با بازیابی (RAG) است تا هم برچسب شایعه بودن/نبودن بدهد و هم تحلیل همراه با توضیح با استناد به منابع بازیابی‌شده ارائه کند.

2) ورودی‌ها و خروجی‌های مدل/سیستم

ورودی سیستم:

• یک متن یا سوال وارد شده توسط کاربر که حاوی ادعایی در مورد سلامت است (مثلاً: "نوشیدن آب اکسیژنه می‌تواند تومورها را درمان کند"). این ورودی در پرامپت «User Input» قرار می‌گیرد.

روجی استاندارد شده:

باید حداقل شامل دو بخش باشد:

1) [Conclusion] : یکی از این سه برچسب می‌باشد

Rumor / Not rumor / Not related to healt information

2) [Analysis] شامل استدلال و تحلیل درباره همان ورودی است.

اگر سیستم از اسناد بازیابی‌شده استفاده کند، در متن تحلیل ارجاع‌هایی مثل [1] می‌آورد و در انتها یک بخش [References] شامل مشخصات منبع اضافه می‌شود (این بخش طبق مقاله در پس‌پردازش تکمیل می‌شود).

3) داده‌های مورد استفاده (نوع، منبع، اندازه)

۳-۱) دیتاست آموزشی: HealthRCN

این دیتاست برای رفع مشکل کمبود داده ساخته شده است.

- **اندازه:** ۱.۱۲ میلیون نمونه داده.
- **منبع اولیه:** (Crawling) پرسش‌های واقعی کاربران از وب‌سایت‌های مشاوره پزشکی معتبر (مانند *net.39*). حدود ۲ میلیون پرسش اولیه جمع‌آوری و پس از پاکسازی به ۱.۱۲ میلیون تقلیل یافت.
- **تولید داده‌های شایعه (Data Augmentation):** از آنجا که سوالات واقعی لزوماً شایعه نیستند، نویسندگان از مدل **GPT-3.5-Turbo** برای تولید نسخهٔ «شایعه» از روی پاسخ‌های صحیح استفاده کردند.
- **ساختار هر نمونه:** شامل پرسش اصلی، عنوان شایعه، متن شایعه (کوتاه/بلند)، پاسخ ردّ شایعه (Debunking)، پاسخ صحیح پزشکی و کلیدواژه‌ها.

۳-۲) پایگاه دانش مرجع (Knowledge Base for Retrieval)

این پایگاه محلی است که سیستم در زمان پاسخ‌دهی (Inference) در آن به جستجوی سند می‌پردازد.

- **منابع:** ۱۹ وب‌سایت و سازمان معتبر سلامت (مانند سایت‌های دولتی بهداشت و بیمارستان‌های تخصصی).
- **اندازه :**
- بخش Elasticsearch (اسناد کامل): ۳۵۸,۱۱۴ سند.
- بخش Milvus (تکه‌های متن): ۱,۳۴۷,۹۷۶ تکه (chunk).
- Chunking: به دلیل محدودیت مدل امبدینگ m3e ، متن مقاله‌ها (به 512 توکن) تکه‌تکه میشود و سپس embedding تولید می‌شود.
- **فرآیند به‌روزرسانی:** این پایگاه به‌صورتی آپدیت می‌شود که اطلاعات جدید پزشکی در دسترس مدل باشد.

3-3) دیتاست ارزیابی (Evaluation Set)

• اندازه: 2500 نمونه

• برچسب‌ها و توزیع: Rumor: 1497؛ Not rumor: 855؛ Not related to health information: 148

• ترکیبی از نمونه‌های HealthRCN + نمونه‌های غیرمرتبط با سلامت (دستی ساخته‌شده) و بدون هم‌پوشانی با دادهٔ فاین‌تیون.

4) روش پیشنهادی مقاله به زبان ساده + شماتیک یا شبه‌کد.

4-1) ایدهٔ کلی HRDE (چهار بخش اصلی)

HRDE چهار ماژول دارد:

- 1) جمع‌آوری و ذخیره اسناد مرجع
- 2) بازیابی اسناد مرجع
- 3) رتبه‌بندی دوباره (Re-ranking)
- 4) تولید پاسخ تشخیص شایعه توسط LLM

شماتیک ساده جریان:

```
User Input
   |
   ▼
(1) Dual Retrieval
   - Keyword search (Elasticsearch/BM25)
   - Semantic search (Milvus + embeddings)
   |
   ▼
(2) Merge Top-N chunks
   |
   ▼
(3) HyDE Re-ranking
   - generate hypothetical doc with a small fine-tuned LLM
   - cosine similarity + threshold
   |
   ▼
(4) Prompting + Large LLM
   - with refs (if any) / without refs (fallback)
   |
   ▼
Output: [Conclusion] + [Analysis] (+ [References] via post-process)
```

4-2) بازیابی دوگانه (Dual Retrieval)

- استخراج کلیدواژه‌ها برای بازیابی مبتنی بر متن (BM25) از پایگاه داده Elasticsearch .
- ورودی کاربر (توسط m3e) برای بازیابی برداری با embedding از پایگاه داده Milvus.
- هدف این است که هم «دقتِ کلیدواژه‌ای» و هم «پوششِ معنایی» را داشته باشند.

به دلیل اینکه خروجی Elasticsearch مقاله کامل است، chunk می‌شود و با chunkهای Milvus ادغام می‌گردد تا Top-N ساخته شود.
در تنظیمات آزمایش، مقاله می‌گوید از Elasticsearch تعداد ۵ سند کامل و از Milvus تعداد ۲۵ chunk بازیابی می‌شود.

4-3) Re-ranking با HyDE (فیلتر کردن اسنادِ واقعاً مفید)
برای اینکه تعداد chunkها به ورودی LLM اصلی قابل‌مدیریت باشد و نویز کم شود:
• یک «سند فرضی» (hypothetical document) با یک مدل کوچک‌تر (Qwen1.5-4B-Chat فاین‌تیون‌شده) تولید می‌کنند.
• شباهت معنایی (cosine similarity) بین هر chunk و سند فرضی محاسبه می‌شود و رتبه‌بندی انجام می‌گیرد؛ سپس Top-K انتخاب می‌شود و یک آستانه شباهت هم برای حذف موارد خیلی نامرتبط اعمال می‌شود.
• در تنظیمات پایه، Top-K نهایی = ۵ قطعه و آستانه شباهت برابر 0.5 است.

4-4) تولید پاسخ (Generation) با پرامپت‌های دوحالته
• دو نوع پرامپت طراحی می‌شود: با رفرنس و بدون رفرنس. هر پرامپت ۴ بخش دارد: توضیح کار، ورودی‌ها، الزامات دقیق، و فرمت خروجی.
• اگر هیچ سند معتبری بازیابی نشود یا همه در Re-ranking حذف شوند، سیستم با پرامپت بدون رفرنس پاسخ می‌دهد (fallback روی دانش داخلی مدل).

5) نتایج اصلی، محدودیت‌ها و ایده‌های ادامه

5-1) نتایج اصلی (کمی + کیفی)
• معیارهای ارزیابی: Accuracy و F1 برای تشخیص برچسب؛ و سه معیار کیفیِ پاسخ (Relevance, Reliability, Richness). همچنین نرخ پاسخ معتبر (Valid Answer Rate) هم ثبت می‌شود.
• برای نمره‌دهی معیارهای کیفی، مقاله از GPT-4-1106-Preview برای ارزیابی با پرامپت‌های مشخص استفاده کرده است.

عملکرد HRDE:

• average accuracy: 91.04%
• F1: 91.58%

طبق آزمایش‌های انجام‌شده، این مدل تخصصی‌شده حدود ۲٪ عملکرد بهتری نسبت به GPT-4 (نسخه Preview-1106) در تشخیص شایعات چینی داشته است.

Ablation (اثر SFT و RAG):
• جدول Ablation نشان می‌دهد ترکیب SFT+RAG بهترین عملکرد را می‌دهد و استفادهٔ تنها از یکی از آن‌ها ایده‌آل نیست (مثلاً RAG بدون فاین‌تیون می‌تواند مدل را گمراه کند).

نمونهٔ کیفی (Case Study):
• در یک مثال، سیستم درست تشخیص می‌دهد «اطلاعات شکر و دیابت» شایعه نیست و برای تحلیل، چند سند مرجع می‌آورد و حتی به دیدگاه WHO دربارهٔ «قند افزوده و سرطان» اشاره می‌کند تا نگرانی کاربر را کاهش دهد.

5-2) محدودیت‌های اعلام‌شده در مقاله
مقاله دو محدودیت مهم را صریح می‌گوید:
• خطا در موضوعات غیر سلامت: به این دلیل که داده‌های آموزش (Fine-tuning) عمدتاً پزشکی است، مدل گاهی ورودی‌های غیر پزشکی کاربران را به اشتباه تفسیر می‌کند.
•گرچه مدل‌های زبانی سرعت استنتاج بالایی دارند (مدل ۴میلیاردی حدود ۶۹۸ توکن/ثانیه و مدل ۱۴میلیاردی حدود ۱۴۳ توکن/ثانیه پردازش می‌کنند)، اما گلوگاه اصلی سیستم، زمان‌بر بودن فرایند بازیابی اسناد (Retrieval) است که کندی کلی سیستم را ایجاد می‌کند.

برنامه آیندهٔ نویسندگان: بهبود دیتاست دستورالعملی و بهینه‌سازی فرایند RAG برای ارتقای عملکرد کلی.

5-3) ایده‌های ادامه (تحلیلی و پیشنهادی بر پایهٔ همین چارچوب)
با توجه به محدودیت‌های خود مقاله و طراحی HRDE، چند مسیر توسعه منطقی است:
• بالانس‌کردن کلاس "Not related to health information": چون خود نویسندگان می‌گویند این نوع نمونه‌ها در آموزش کم بوده، اضافه کردن نمونه‌های متنوعِ غیرسلامت (و حتی ورودی‌های مرزی مثل تغذیه عمومی/ورزش/سبک زندگی) می‌تواند خطای تفسیر را کم کند.
• کاهش زمان بازیابی: چون سرعت پاسخ‌دهی تحت تأثیر retrieval است، می‌توان روی کاهش تعداد candidates، کش‌کردن نتایج پرتکرار، یا ساده‌سازی rerank در بعضی سناریوها کار کرد (مقاله خودش «زمان‌بر بودن retrieval» را محدودیت می‌داند).