

# 实习工作总结

 ByteDance 字节跳动

AI-Lab-NLP 内容质量项目  
-- 何逸宸

- 伪科学内容识别
  - 信源库构建
  - 通用多模态审核模型
- 小模型训练&部署
  - 工程开发
  - 大模型训练&部署

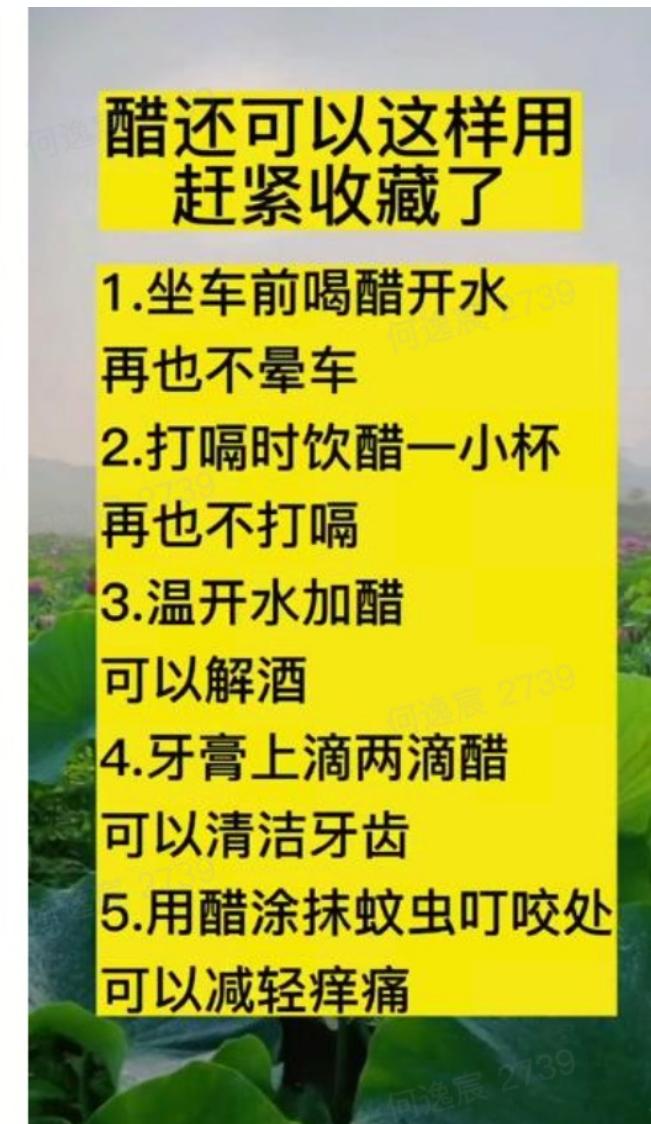
# 内容质量项目

虚假信息的发布和传播会对社会造成很大的危害

- 影响对重大事件的认知 --虚假新闻

## 俄乌冲突

- 俄罗斯国防部称，波兰军队已经进入到了乌克兰，并且正式加入了对俄罗斯的战争，俄罗斯军队表示，不会让波兰军队走掉，如果有必要的话，俄罗斯军队将立即对波兰发动闪电攻击·最新消息
- 12000美军进入，炸毁俄罗斯石油管道！普京下令紧急出兵....
- 俄军开火立陶宛，48枚导弹凌空爆炸！这就是惹怒普京的代价
- 威胁健康，造成财产损失 – 伪科学、知识库（谣言库）**



# 伪科学内容识别

# 伪科学内容识别 文本二分类

伪科学谣言为什么会被传播

内容披着“科学”外衣，具有更强的迷惑性

网络平台提供迅速爆炸性传播的渠道

读者出于“宁可信其有，不可信其无”的避害思维，很少刨根问底、追问究竟

# 保持科学理性的思考方式

识别科学谣言的意义

减少伪科学内容对个人带来的健康和经济利益的损失

不做谣言的传播者，对有害内容进行阻击，维护健康的网络环境

扩大优质内容的运营、传播、推广等动作，同时在谣言治理，危险账号拦截等方面有一些针对性的对策

## “去氘水修复DNA”

清清字泉  
2022-5-11 15:18 · 来自广东

当人体新陈代谢功能受损时，适当饮用去氘水，可对DNA有很好修复的作用。去氘水——是运用现代科学技术，把自然界中的水去除部分氘后得到的水，也加贫氘水，低氘水，超之水。科学研发现，饮用水的氘含量越低，小分子活性越强，容易进入细胞，被生物体吸收利用，对保持身体健康具有重要意义。



1336展现

评论

1 赞 | 0 转发

暂无评论，点击抢沙发

## “院士只会讲脱口秀”

彗星撞火星  
2022-1-2 08:34

霍金早就警告过人类不要接触外星人，但某位国内的学者硬要公开演讲说要试图联系外星人？这位国内学者兴奋说道国外最近探测到了外星人的信号源，并且表示要快点建立连接。。。 (无语) 怪不得出不了一个诺贝尔，别人的科学院士拿诺贝尔，国内的院士讲脱口秀，还不是自己发现并研究的。也不管霍金说过什么！这难道就是差距吗？#我要上头条#

5748展现

评论 2 4 赞 | 0 转发

1 赞 | 0 转发

但凡能来地球我们就直接投降好了

回复 > 1-2

差距太大

回复 > 1-2

暂无评论，点击抢沙发

何逸辰 2739

# 伪科学内容识别 文本二分类

## 写作风格

教育教学

写真地理

2020年6月 第22期

### 熟鸡蛋变成生鸡蛋(鸡蛋返生)——孵化雏鸡的实验报告

郭平 白卫云  
(郑州市春霖职业培训学校 河南 郑州 450000)

摘要：“鸡蛋返生”，顾名思义，就是由熟鸡蛋再变成生鸡蛋。这是一个难以想象的，甚至是不可逆的，但是这样神奇的现象确实在郑州春霖职业培训学校发生了。一群特别培训的学生，在郭平老师的指导下，正在进行一个神奇实验，即熟鸡蛋重新变成生鸡蛋，并将返生后的生鸡蛋进行孵化成雏鸡。并且已经成功返生了40多枚。

关键词：生鸡蛋；熟鸡蛋；鸡蛋返生；孵化雏鸡；实验报告  
【中图分类号】S831 【文献标识码】A

鸡蛋奇特返生的现象，根据鸡蛋的组织结构及功能。鸡蛋经过高温100℃开水煮20分钟，变成熟鸡蛋后，学生们运用自己的超心理意识能量方法等，将这些熟鸡蛋变成生鸡蛋，现在我们将这种奇特现象分享给科学探索爱好者，共同探究其内在理论依据。

实验材料：鸡蛋10枚，一次性纸质茶杯10个。

实验场地：郑州春霖学校507教室。

实验时间：2020年8月12日11时。

室内温度：摄氏25℃，保持室内安静。

参加人员：郑州春霖学校特训生10人。见表1。

表1 观察见证专家及学生家长

姓名	职务(职称)	单位
郭萍	校长	郑州市春霖职业培训学校
郭大勇	教授	兰州毕业大学书记
马建民	院长	兰州大学设计院系分院院长
单松柏	原国家地震局兰州地质中心主任	革质菌家津贴
白卫云	主任医师	河南医学高等专科学校附属医院
尹杰		山东理工大学讲师
高爱华	院长	晋城县普明中学
孙静霞	院长	孙嘉金海小学
赵真静	院长	晋都里小学
郭大勇	教授	新郑市普明小学副校长



【文章编号】1674-3733(2020)22-0224-01

钙等矿物质组成，对鸡蛋内容物起保护作用。②外层卵壳膜，是层无结构纤维膜，主要是保护鸡蛋内容物水分不丢失。③内层卵壳膜，是一种可透气膜，空气可以进出。④气室，位于鸡蛋的钝端内(在大头与卵壳膜之间)，是由两层卵壳膜之间常分开形成一个小气室，贮存空气，具有胚胎发育时供应其呼吸功能<sup>[2]</sup>。⑤系带，卵黄的两端由浓郁的蛋白质组成卵黄系带，其功能是维持卵细胞共定于蛋白中心位置，对卵细胞具有起着缓冲作用，可防止卵细胞的震荡。⑥卵黄膜，位于卵白与卵黄之间的一层薄膜，是卵细胞的组成部分。⑦卵黄：呈黄色，位于细胞的中央，是鸡卵胚胎发育的主要营养物质。⑧胚盘：位于卵黄表面中央有一圆形盘状小白点，呈椭球状等，卵白：卵壳膜与卵黄膜之间，为胚胎发育提供水和营养物质，卵黄表面中央有一圆盘状的小白点(就是在蛋黄上看到的小白点)称为胚盘，里面含有细胞核，未受精的卵，胚盘色淡而小，已受精的卵，胚盘色浓而略大，这是因为胚胎发育已经开始<sup>[2]</sup>。如果是受精卵，胚盘在适宜的条件下就能孵化出雏鸡，胚盘进行胚胎发育的部位<sup>[2]-[3]</sup>。

我们知道，蛋白质加热后可以变性，那么，熟鸡蛋经过100℃开水煮20分钟，整体上鸡蛋内容物均有液态变成固态，在返生过程中，不添加任何化学物质，不进行任何物理处理，如加温或者降温、电离辐射等。这是为什么？

鸡卵细胞的由卵黄膜、卵黄和胚盘等组成，卵细胞是否变性，如果变性，即使返生为液态，也难以孵化成雏鸡。这是为什么？欢迎讨论问题如下：

## 文本、字体、排版特征

但是应该很少有人知道，身为科学家的焦耳其实信奉神。

事实上，很多大名鼎鼎的“老”英国科学家都信神，比如近代物理学之父牛顿、统计物理学的奠基人之一麦克斯韦、电学之父和交流电之父法拉第、热力学之父开尔文勋爵……



好吧，超模君还是承认好了，实在不知道怎么写开头，就给大家科普了上面这个小知识。

其他人都拉来活跃气氛的，只有詹姆斯·普雷斯科特·焦耳（James Prescott Joule）才是今天的主人公。

### 酿酒小子

ByteDance 字节跳动

除了信奉神，焦耳身上还有两大标签：物理学家和英国皇家学会会员。

# 伪科学内容识别 文本二分类

## 写作风格

教育教学

写真地理

2020年6月 第22期

### 熟鸡蛋变成生鸡蛋(鸡蛋返生)——孵化雏鸡的实验报告

郭平 白卫云

(郑州市春霖职业培训学校 河南 郑州 450000)

**摘要：**“鸡蛋返生”，顾名思义，就是由熟鸡蛋再变成生鸡蛋。这是一个难以想象的，甚至是不可逆的，但是这样神奇的现象确实发生在郑州春霖职业培训学校发生了。一群特别培训的学生，在郭平老师的指导下，正在进行一个神奇实验，即熟鸡蛋重新变成生鸡蛋，并将返生后的生鸡蛋进行孵化成雏鸡。并且已经成功返生了40多枚。

**关键词：**生鸡蛋；熟鸡蛋；鸡蛋返生；孵化雏鸡；实验报告  
【中图分类号】S831 【文献标识码】A

鸡蛋奇特返生的现象，根据鸡蛋的组织结构及功能。鸡蛋经过高温100℃开水煮20分钟，变成熟鸡蛋后，学生们运用自己的超心理意识能量方法等，将这些熟鸡蛋变成生鸡蛋，现在我们将这种奇特现象分享给科学探索爱好者，共同探究其内在理论依据。

**实验材料：**鸡蛋10枚，一次性纸质茶杯10个。

**实验场地：**郑州春霖学校507教室。

**实验时间：**2020年8月12日11时。

**室内温度：**摄氏25℃，保持室内安静。

**参加人员：**郑州春霖学校特训生10人。见表1。

表1 见章见证专家及学生家长

姓名	职务(职称)	单位
郭萍	校长	郑州市春霖职业培训学校
宋大勇	教授	兰州毕业大学书记
马建民	院长	兰州大学设计院系分院院长
单松柏	原国家地震局兰州地质中心主任	革质菌家津贴
白玉忠	主任医师	河南医学院高等专科学校附属医院
尹杰		山东医科大学
高爱华	院长	贵州遵义医学院(三届临床师)
孙静霞	院长	孙嘉金师
赵真静	院长	贵州医大
郝大勇	教授	新郑市首德小学副校长



**【文章编号】**1674-3733(2020)22-0224-01

鸡蛋矿物质组成，对鸡蛋内容物起保护作用。①外层卵壳膜，是层无结构纤维膜，主要是保护鸡蛋内容物水分不丢失。②内层卵壳膜，是一种可透气膜，空气可以进出。③气室，位于鸡蛋的钝端内(在大头与卵壳膜之间)，是由两层卵壳膜之间常分开形成一个小气室，贮存空气，具有胚胎发育时供应其呼吸功能<sup>[1]</sup>。④系带：卵黄的两端由浓郁的蛋白质组成卵黄系带，其功能是维持卵细胞共定于蛋白中心位置，对卵细胞具有起着缓冲作用，可防止卵细胞的震荡。⑤卵黄膜：位于卵白与卵黄之间的一层薄膜，是卵细胞的组成部分。⑥卵黄：呈黄色，位于细胞的中央，是鸡卵胚胎发育的主要营养物质。⑦胚盘：位于卵黄表面中央有一圆形盘状小白点，呈椭球状等，卵白：卵壳膜与卵黄膜之间，为胚胎发育提供水和营养物质，卵黄表面中央有一圆盘状的小白点(就是在蛋黄上看到的小白点)称为胚盘，里面含有细胞核，未受精的卵，胚盘色淡而小，已受精的卵，胚盘色浓而略大，这是因为胚胎发育已经开始<sup>[2]</sup>。如果是受精卵，胚盘在适宜的条件下就能孵化出雏鸡，胚盘进行胚胎发育的部位<sup>[2-3]</sup>。

我们知道，蛋白质加热后可以变性，那么，熟鸡蛋经过100℃开水煮20分钟，整体上鸡蛋内容物均有液态变成固态，在返生过程中，不添加任何化学物质，不进行任何物理处理，如加温或者降温、电离辐射等。这是为什么？

鸡卵细胞的由卵黄膜、卵黄和胚盘等组成，卵细胞是否变性，如果变性，即使返生为液态，也难以孵化成雏鸡。这是为什么？欢迎讨论问题如下：

## 文本、字体、排版特征

但是应该很少有人知道，身为科学家的焦耳其实信奉神。

事实上，很多大名鼎鼎的“老”英国科学家都信神，比如近代物理学之父牛顿、统计物理学的奠基人之一麦克斯韦、电学之父和交流电之父法拉第、热力学之父开尔文勋爵

但是应该很少有人知道，**身为科学家的焦耳其实信奉神**。事实上，很多大名鼎鼎的“老”英国科学家都信神，比如近代物理学之父牛顿、统计物理学的奠基人之一麦克斯韦、电学之父和交流电之父法拉第、热力学之父开尔文勋爵……**好吧，超模君还是承认好了，实在不知道怎么写开头，就给大家科普了上面这个小知识。其他人都是拉来活跃气氛的，只有詹姆斯·普雷斯科特·焦耳 (James Prescott Joule) 才是今天的主人公。****酿酒小子**除了信奉神，焦耳身上还有两大标签：物理学家和英国皇家学会会员。

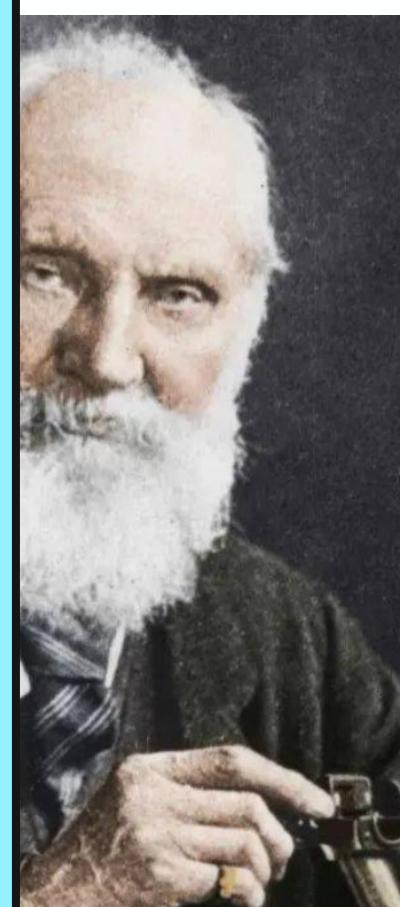
好吧，超模君还是承认好了，实在不知道怎么写开头，就给大家科普了上面这个小知识。

其他人都是拉来活跃气氛的，只有詹姆斯·普雷斯科特·焦耳 (James Prescott Joule) 才是今天的主人公。

## 酿酒小子

ByteDance 字节跳动

除了信奉神，焦耳身上还有两大标签：物理学家和英国皇家学会会员。



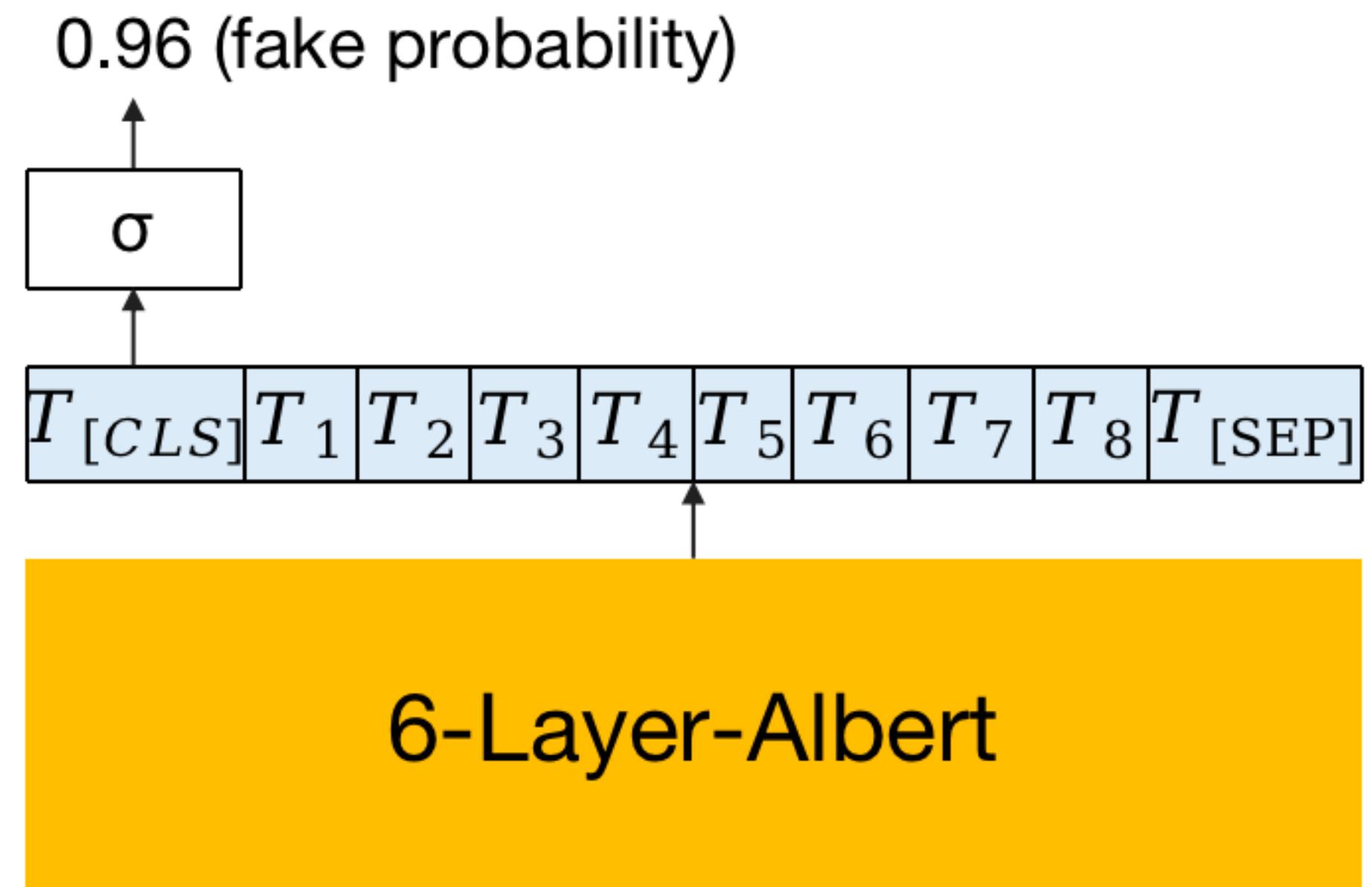
# 伪科学内容识别 文本二分类

## 数据来源

- 头条：图文、微头条、问答
- 抖音：小视频标题以及通过ocr和asr获取的文字信息
- 西瓜：中视频标题以及通过ocr和asr获取的文字信息

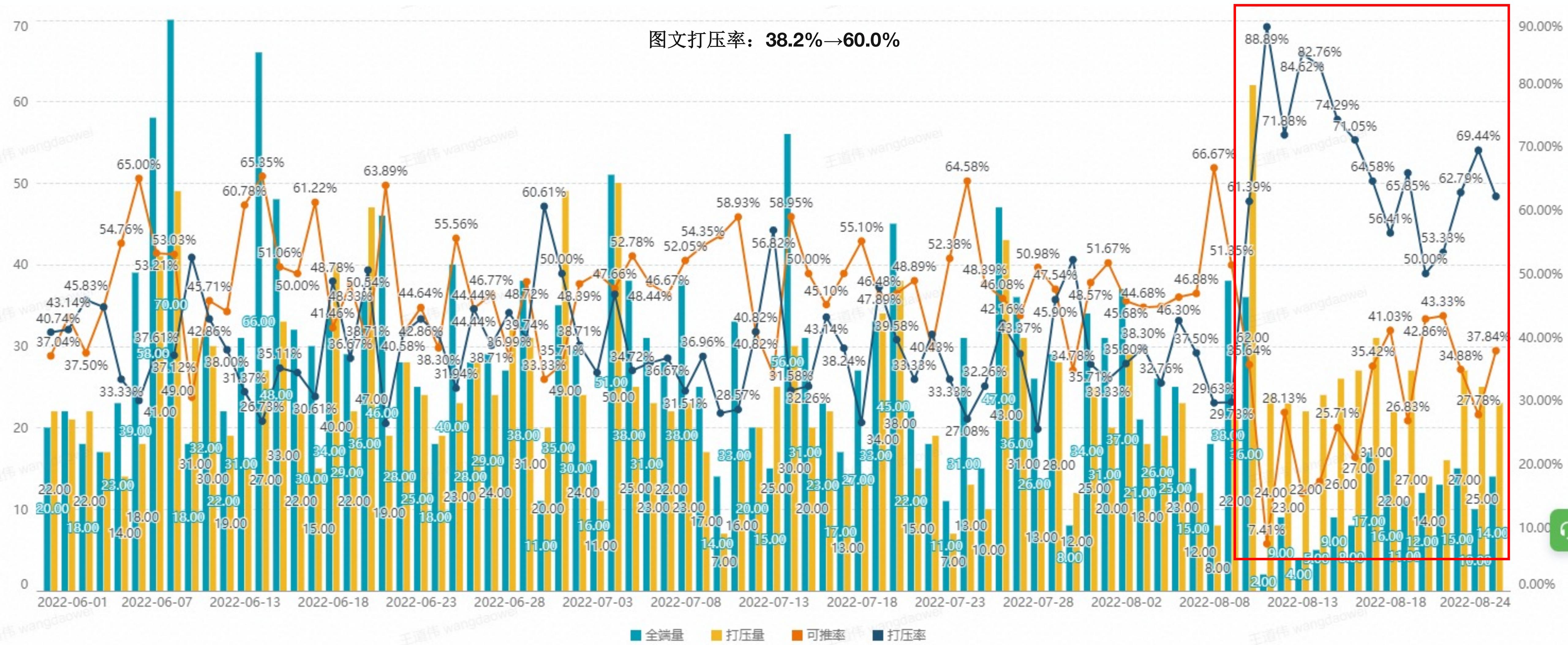
## 现有模型的更新

- 数据抓取和清洗：使用SQL在数据库中抓取数据，筛选出人工审核后进行打压的数据作为正例，不打压的数据作为负例。
- 数据增强：由于正例过少，采用正利double，负例亚采样的方式，保证正负例比例在1:5左右。
- 构造数据集：按时间由远及近分别构造训练集、验证集和测试集，比例大约为10:1:1。
- 模型训练及验证：训练模型，使用验证集选取checkpoint并在测试集上进行测试。
- 人工检验模型效果：将模型命中的数据进行人工标注，验证模型的打压率。

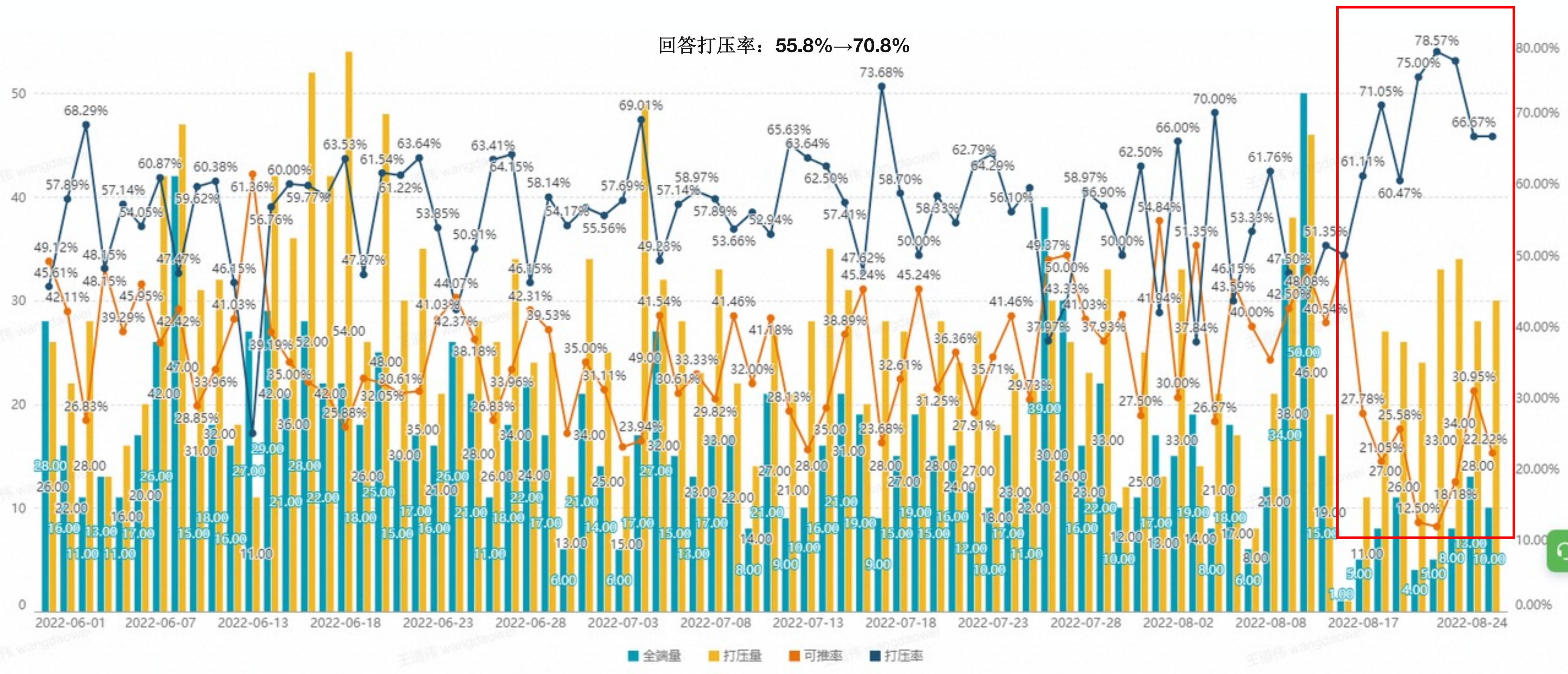


体裁	线上打压滤	新模型打压率
图文	40%	48.5%
微头条	24%	36.6%
回答	54%	57%
小视频	18.4%	50.9%
中视频	--	31.7%

# 伪科学内容识别 文本二分类



# 伪科学内容识别 文本二分类



# 伪科学内容识别 文本二分类

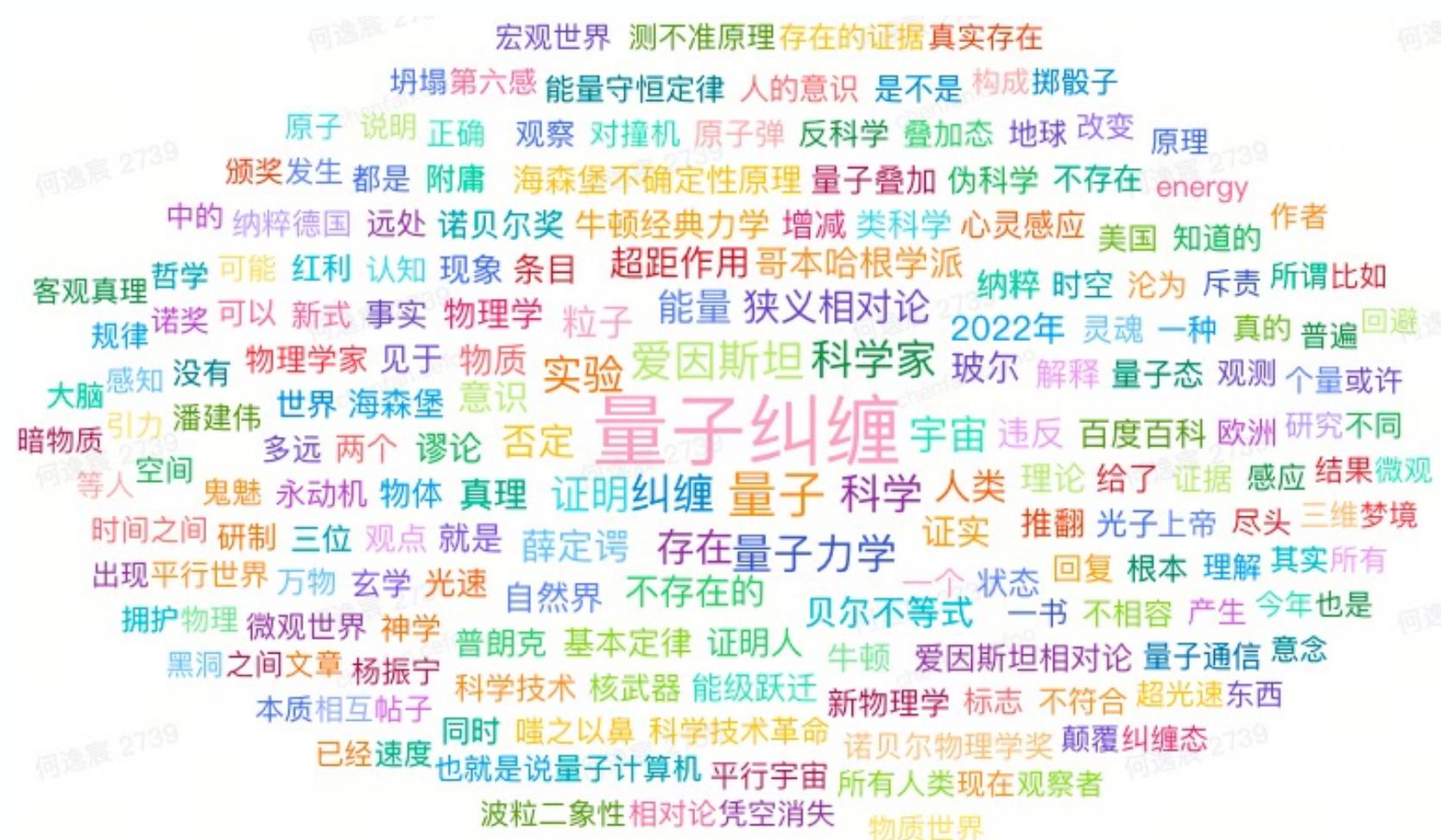
## 结果和收益

1. 对于四体裁伪科学模型进行模型训练，打压率由30%提升至40%
2. 接入了新的中视频体裁，丰富了项目拦截范围，目前中视频周审出1000+，打压率超40%
3. 伪科学周均拦截科学虚假内容700+，年度拦截伪科学虚假内容3.5W+
4. 对全局内容的调节，全局科学虚假内容占比图文三体裁从2022年3月的6.2%虚假占比，降低到了3.8%

## 未来改进计划：实现模型自动更新

模型上线一段时间后，作者会根据被打压的情况改变写作风格，导致模型的准确率下降。因此每隔一段时间需要重新抓取数据进行训练。

整个训练测试上线流程较固定，可实现自动化流程。自动抓取数据，进行数据增强，训练模型，在相同的测试集中测试新旧模型的效果，效果有提升则自动更新线上模型。



波顿vlog 原创

2022-9-9 05:55 · 来自浙江

关注

此内容尚未得到可靠来源证实，请注意甄别内容真实性

其实这个话题，目前来说也没有哪个科学家能给出一个确切的结论说金字塔到底是不是外星人建的，但是也不能排除这种可能。即使到现在金字塔还是有很多的未解之谜，即使用现在的技术去

# 信源库系统搭建

# 信源库系统搭建 项目背景

## 白名单-信源库

权威信源库：覆盖国内外权威信源发布信息。

- 头条站内规范稿源
- 权威媒体（站点150+）
  - 新华社 / 光明网 / 央视网 / ...
  - 政府机构 / 公司官网 / ...
  - 垂类权威媒体
  - 路透社 / CNN / 法新社 / 半岛电视台 / 俄罗斯卫星通讯社 / ...
  - 收录文章数400w+, 日均新增5k+

## 黑名单-谣言库

包含谣言3w+, 周新增200+

- 胡萝卜能抗衰老
- 凯里剑河县交警收黑钱
- 雷锋被从教材中删减
- 山楂能降血脂

蕴含谣言，虚假！

[文章]: 胡萝卜又称小人参，没想到它的好处这么多！不仅可以补肝明目，降糖降脂，还能增强免疫力和抗衰老！

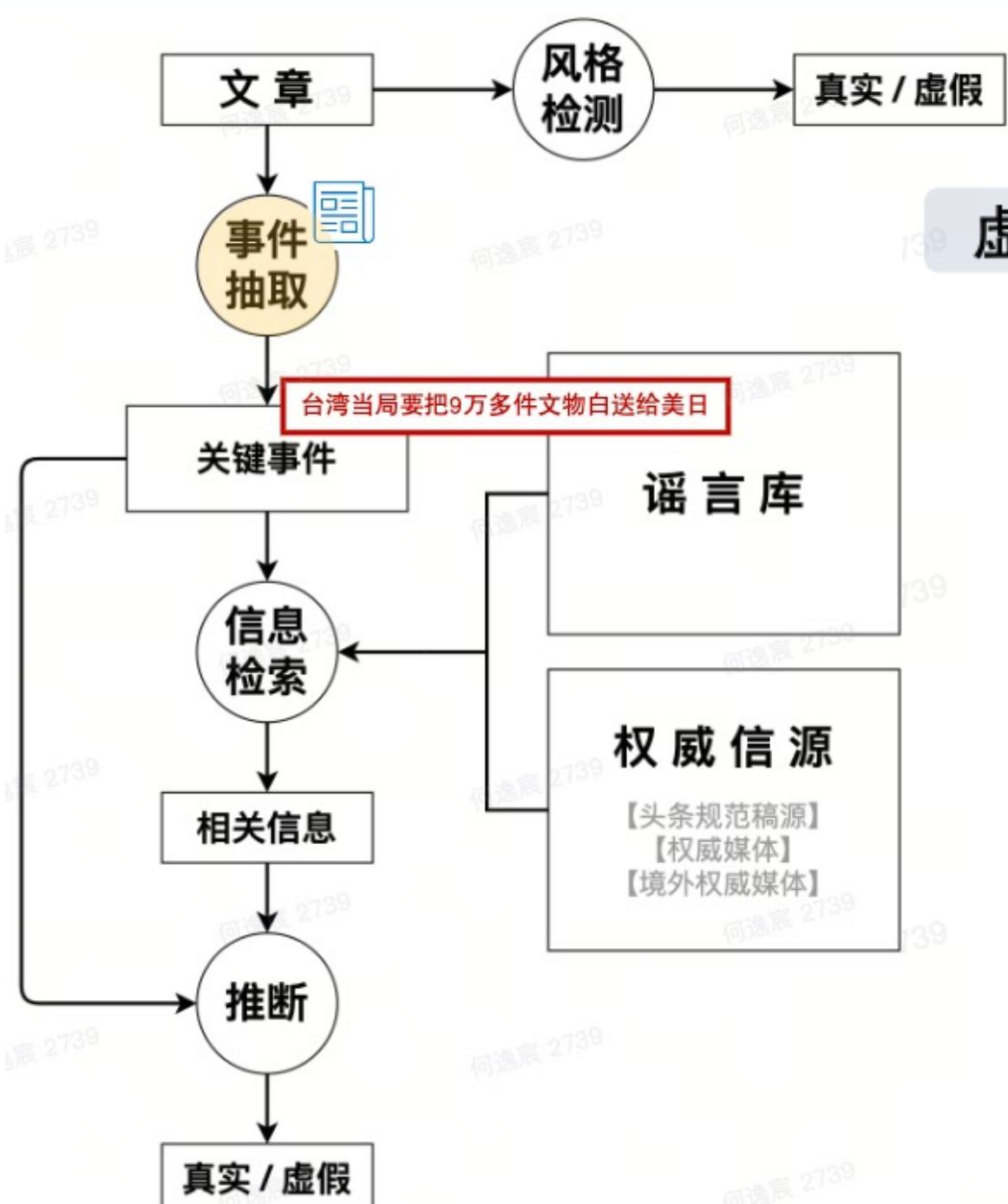


最近台湾局势非常的紧张，台湾方面已经乱了，最近有一个来自海外网的消息称：**台湾当局要把9万多件文物白送给美日**！要知道每件文物的价值都在9万美金以上，有些甚至要数千万数亿万的国宝。

表面上这是一次紧急的疏散的演练，但是实际是在为这些文物想到转移方案，自从佩洛西来台湾后，台湾的局势发生了很大的变化，一旦发生不测，这些文物就有可能落到大陆的手中。

但是关于这批文物的细节和内容，相关人员却一直没有说。

2022-08-08 17:16:57



虚 假

台湾当局要把9万多件文物白送给美日, 2022-08-08 17:16:57

**南方plus客户端**: 台北故宫辟谣台当局准备将文物转移至美日：“绝无此事”， 2022-08-08 17:14:00

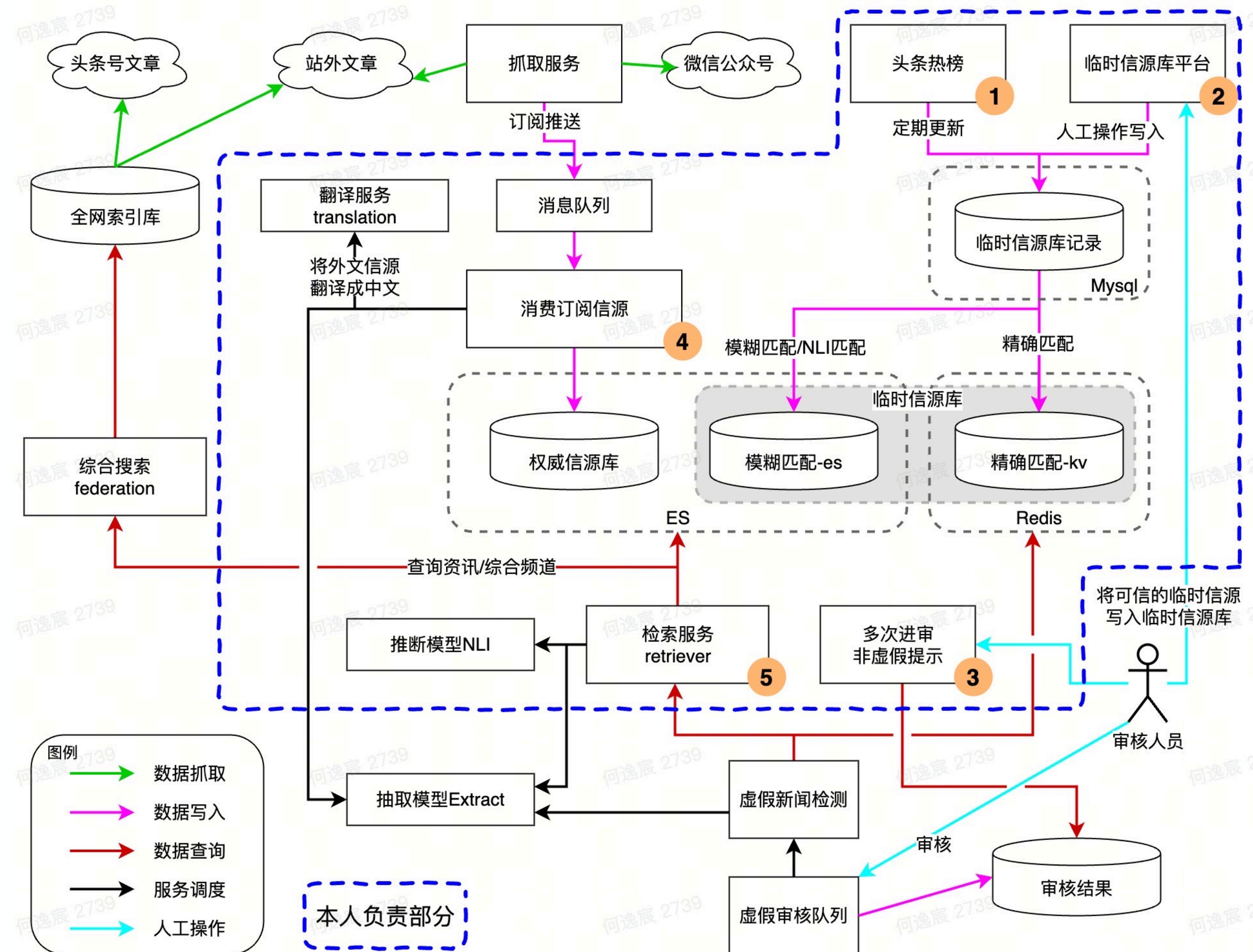
- **澎湃新闻**: 岛内传台当局要将故宫文物送美日“保护”，台北故宫方面支吾，2022-08-08 16:11:24
- **海峡网**: 传民进党当局准备将9万藏品转移美日寻求“保护”，台北故宫回应了，2022-08-08 15:04:39
- .....

### 虚假模型信息

文中短句1:	台湾当局要把9万多件文物白送给美日
判断结果:	与权威信源矛盾，疑似虚假
命中权威信源信息:	台北故宫辟谣台当局准备将文物转移至美日：“绝无此事”
南方plus客户端	2022-08-08 17:14:00
岛内传台当局要将故宫文物送美日“保护”、台北故宫方面支吾	澎湃新闻 2022-08-08 16:11:24
9万藏品将转移美日？台北故宫回应	鲁网 2022-08-08 16:08:30
传民进党当局准备将9万藏品转移美日寻求“保护”、台北故宫回应了	海峡网 2022-08-08 15:04:39
台北故宫博物院回应：绝无此事	西安晚报 2022-08-08 13:43:28

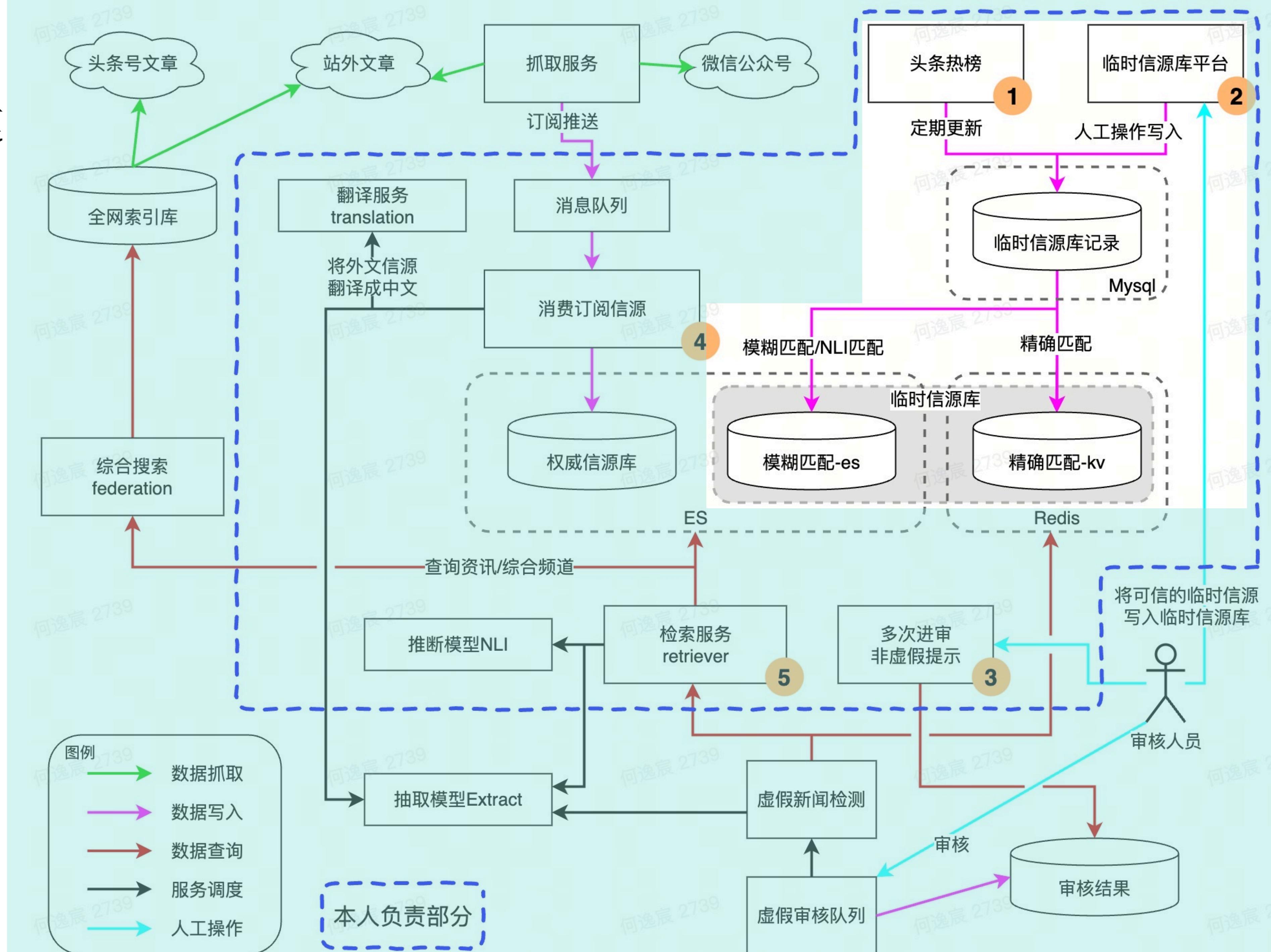
# 信源库系统搭建

1. 热榜内容自动入库
2. 临时信源库
3. 多次进审非虚假提示
4. 订阅信源消费
5. 检索服务



# 信源库系统搭建

1. 热榜内容自动入库
2. 临时信源库
3. 多次进审非虚假提示
4. 订阅信源消费
5. 检索服务



# 热榜内容自动入库 Python | Redis | MySQL

微头条有许多发文是针对热榜事件的评论，热榜内容很多是突发的高热度事件，其内容是经过审核的但是没有录入权威信源库，因此会造成大量的没有权威信源的情况。



运筹帷幄荷叶nn

2022-08-21 11:15

+ 关注

#赖岳谦评台军与解放军军舰对峙画面# 赖教授说的对！为他点赞！

虚假模型信息

文中短句1：赖岳谦评台军与解放军军舰对峙画面

判断结果：无相关权威信源，疑似虚假

命中权威源 无

信息：

① 举报

使用定时job，每隔5min抓取一次热榜内容，并将其加入临时信源库**Abase**中，每次加入的内容设置其**ttl**为30min。模型抽取出的内容会与临时信源库进行nli比对，若命中则不进审，日均减少进审量约100条。

```
client = Client(psm="", table="")
url = "https://api.toutiaoapi.com/hot-event/hot-board/?origin=hot_board"
create_time = int(datetime.now().timestamp()) # 获取当前时间戳
response = requests.get(url) # 发送 HTTP GET 请求
data = json.loads(response.text)["data"] # 解析 JSON 响应
db = connect(db_psm="")
cursor = db.cursor()
for cluster in data:
    title = cluster["Title"].strip() # 提取热榜标题
    client.set(title, "true", ex=1800) # 将标题写入 Bytedatabase
    cursor.execute("INSERT INTO temp_trusted_data(statement, create_time, match_method, ttl, owner, status VALUES (%s, %s, %s, %s, %s, %s) on duplicate key update ttl = %s", (title, create_time, 1, create_time + 1800, "hot_board", create_time + 1800, 1))
db.commit() # 提交 MySQL 事务
cursor.close() # 关闭数据库连接和游标
db.close()
```

# 临时信源库 Golang | Redis | MySQL | ES

存在许多类似热榜的临时信源，因为热度不够高没能进入热榜，我们将一些多次进审不打压的case视为这种临时信源通过飞书bot反馈给运营同学，在他们人工判断的情况下决定是否将其加入临时信源库(MySQL, 用于前端展示)，并设置精确匹配(Redis)和模糊匹配(ES)。

本人主要负责临时信源库后端开发，使用Golang语言的Hertz框架，支持临时信源的增删改查功能，方便运营同学查看和编辑。

```
func customizedRegister(r *server.Hertz) {
    r.POST("/add", handler.Add)           // 增
    r.DELETE("/delete", handler.Delete)   // 删
    r.PUT("/update", handler.Update)     // 改
    r.GET("/query", handler.Query)       // 查
}
```

The screenshot shows a web-based management interface for a temporary source library. At the top right is a 'Create' dialog box with fields for 'Event Content' (输入临时新源内容), 'Creator' (创建人), and 'Expiration Time (Default one day valid)' (过期时间 (默认一天有效)). The 'Exact Match' (精确匹配) tab is selected. Below the dialog is a table listing existing temporary sources. The table has columns: ID, Event, Matching Mode, Status, Expiration Time, Operator, Last Update Time, and Operations (Edit, Archive). A red arrow points from the 'Create' dialog's 'Event Content' field to the 'Event' column of the table. Another red arrow points from the 'Edit' button in the table's operations column to the 'Edit' button in the 'Create' dialog.

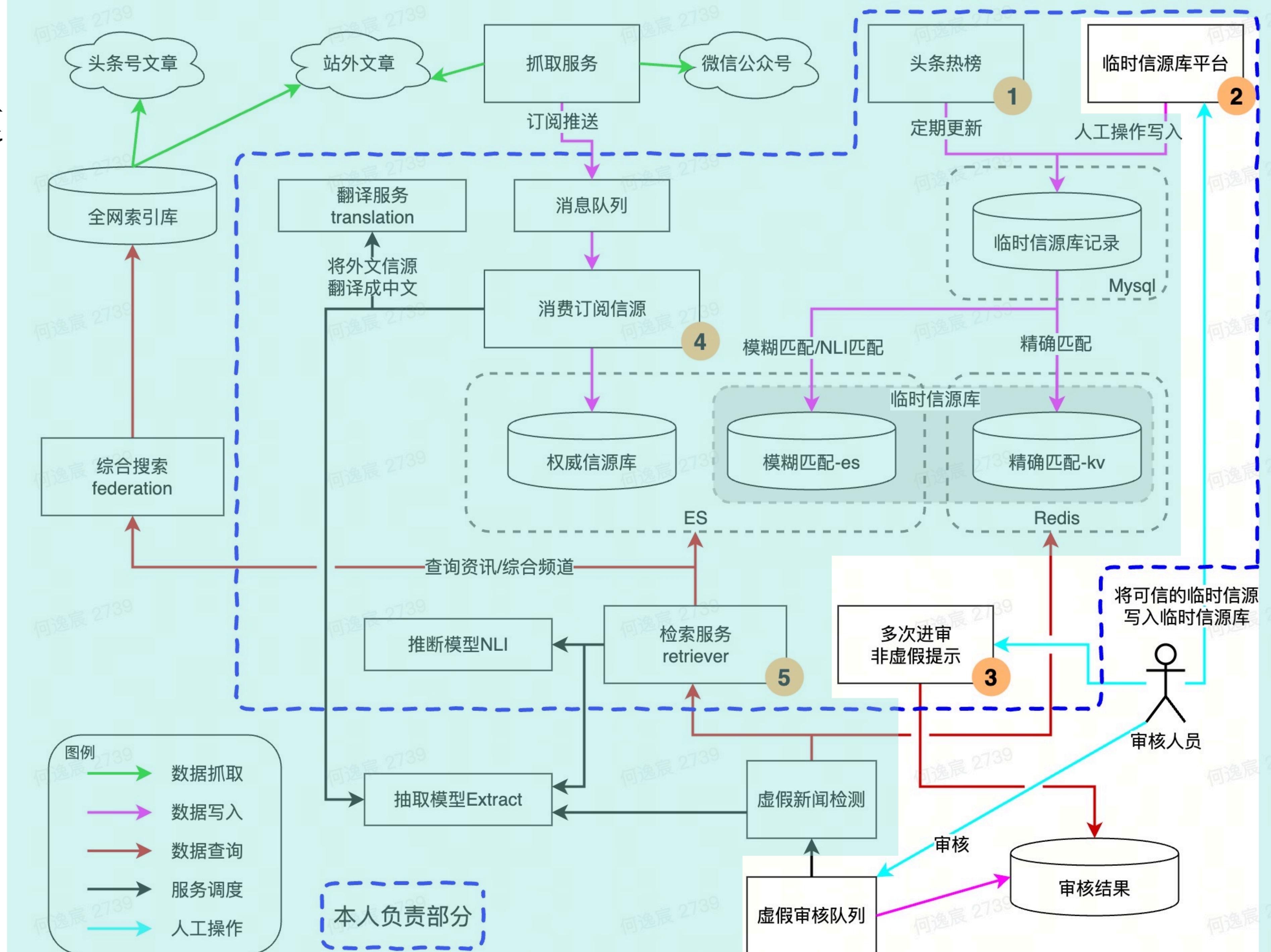
ID	事件	匹配模式	状态	过期时间	操作人	最后更新时间	操作
38481	中国电科CETC加班事件	NLI匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/6 09:30:30	<button>编辑</button> <button>下架</button>
38479	美国政府周三表示，他们已对格鲁吉亚最高司法委员会的4名成员实施签证制裁，理由是他们涉嫌“严重腐败”	NLI匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/6 09:22:47	<button>编辑</button> <button>下架</button>
38457	河北8次地震了	精确匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/6 07:38:30	<button>编辑</button> <button>下架</button>
38314	五角大楼的一份报告，中国首次将至少一艘核武装弹道导弹潜艇保持在海上	NLI匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/5 14:06:23	<button>编辑</button> <button>下架</button>
38310	西南大学一教授被曝潜规则女博士三年	NLI匹配	过期	2023/4/6 00:00:00	luojingshu	2023/4/5 13:53:03	<button>编辑</button> <button>下架</button>

共 221 条 < 1 2 3 ... 23 >

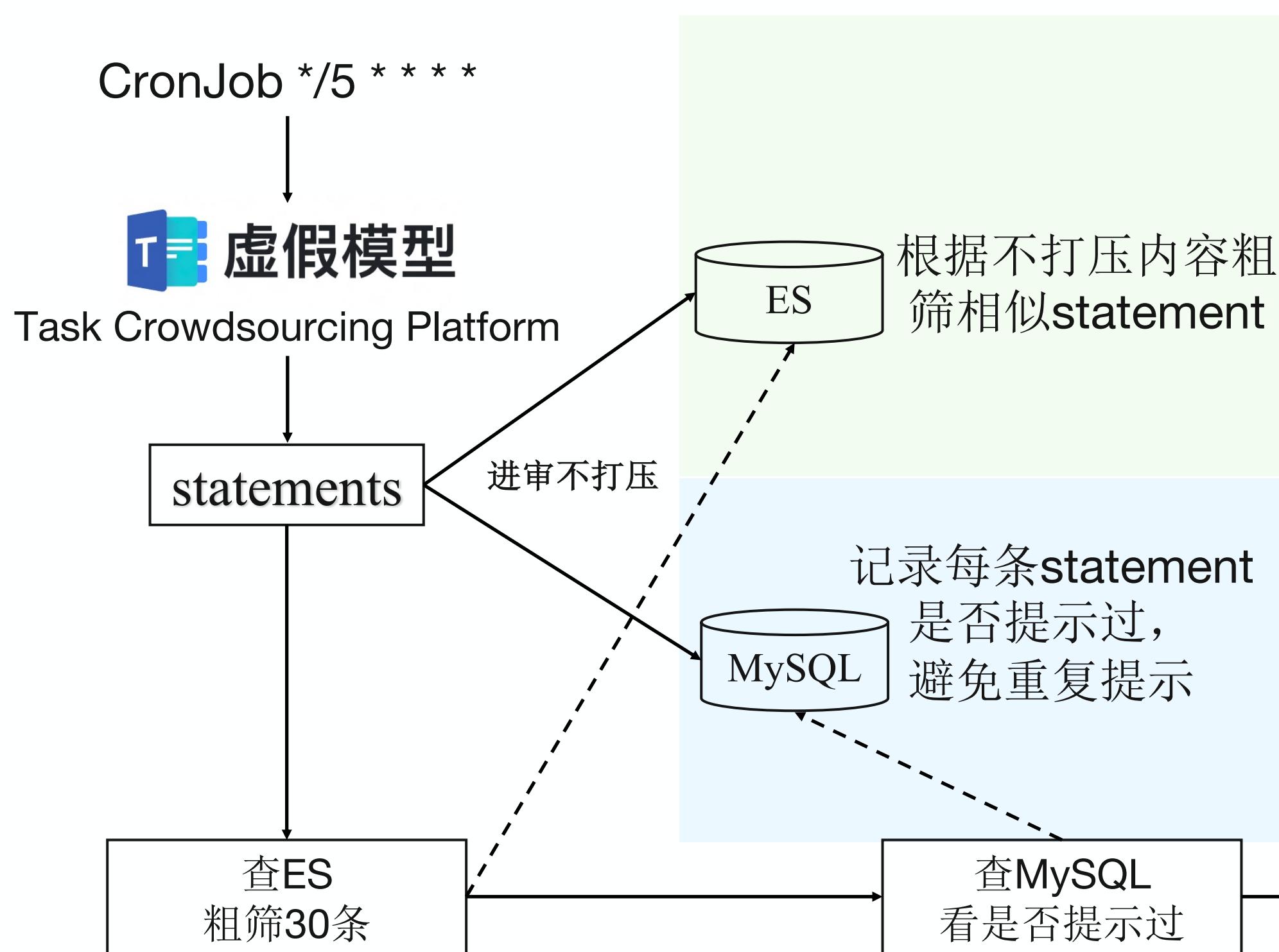
字节跳动

# 信源库系统搭建

1. 热榜内容自动入库
2. 临时信源库
3. 多次进审非虚假提示
4. 订阅信源消费
5. 检索服务

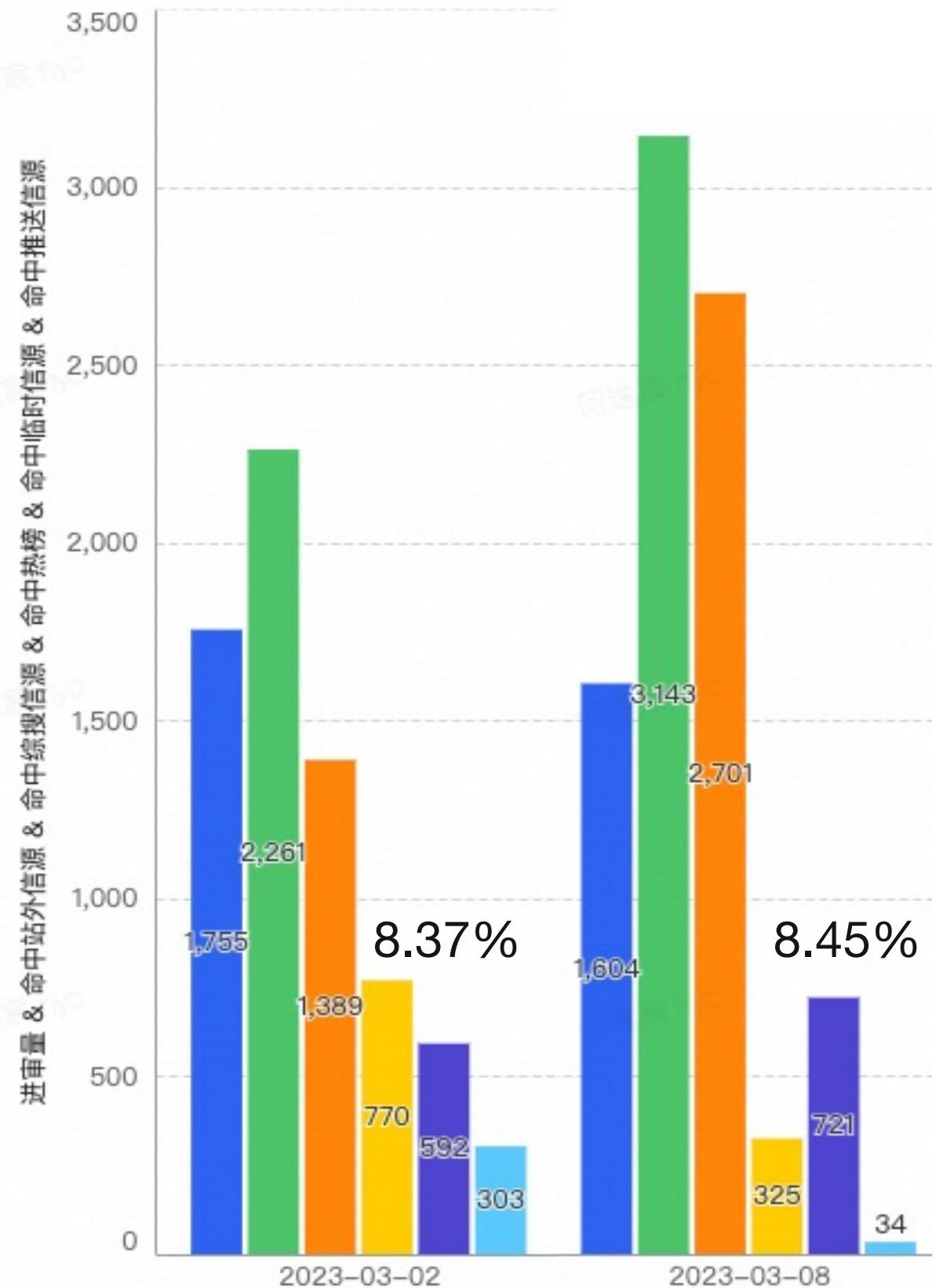


# 多次进审非虚假提示 Python | MySQL | ES | CronJob



```
"gid": {  
    "type": "keyword"  
},  
"statement": {  
    "fields": {  
        "keyword": {  
            "ignore_above": 256,  
            "type": "keyword"  
        }  
    },  
    "type": "text"  
}
```

```
CREATE TABLE `noticed_data` (  
    `id` bigint unsigned NOT NULL AUTO_INCREMENT COMMENT '自增ID',  
    `statement` varchar(255) NOT NULL COMMENT '事件描述',  
    `notice_time` int DEFAULT NULL COMMENT '提醒时间',  
    `status` int DEFAULT '1' COMMENT '是否打压',  
    `gid` varchar(64) NOT NULL COMMENT '文章id',  
    PRIMARY KEY (`id`),  
    UNIQUE KEY `uniq_statement_gid` (`statement`,`gid`)  
)
```

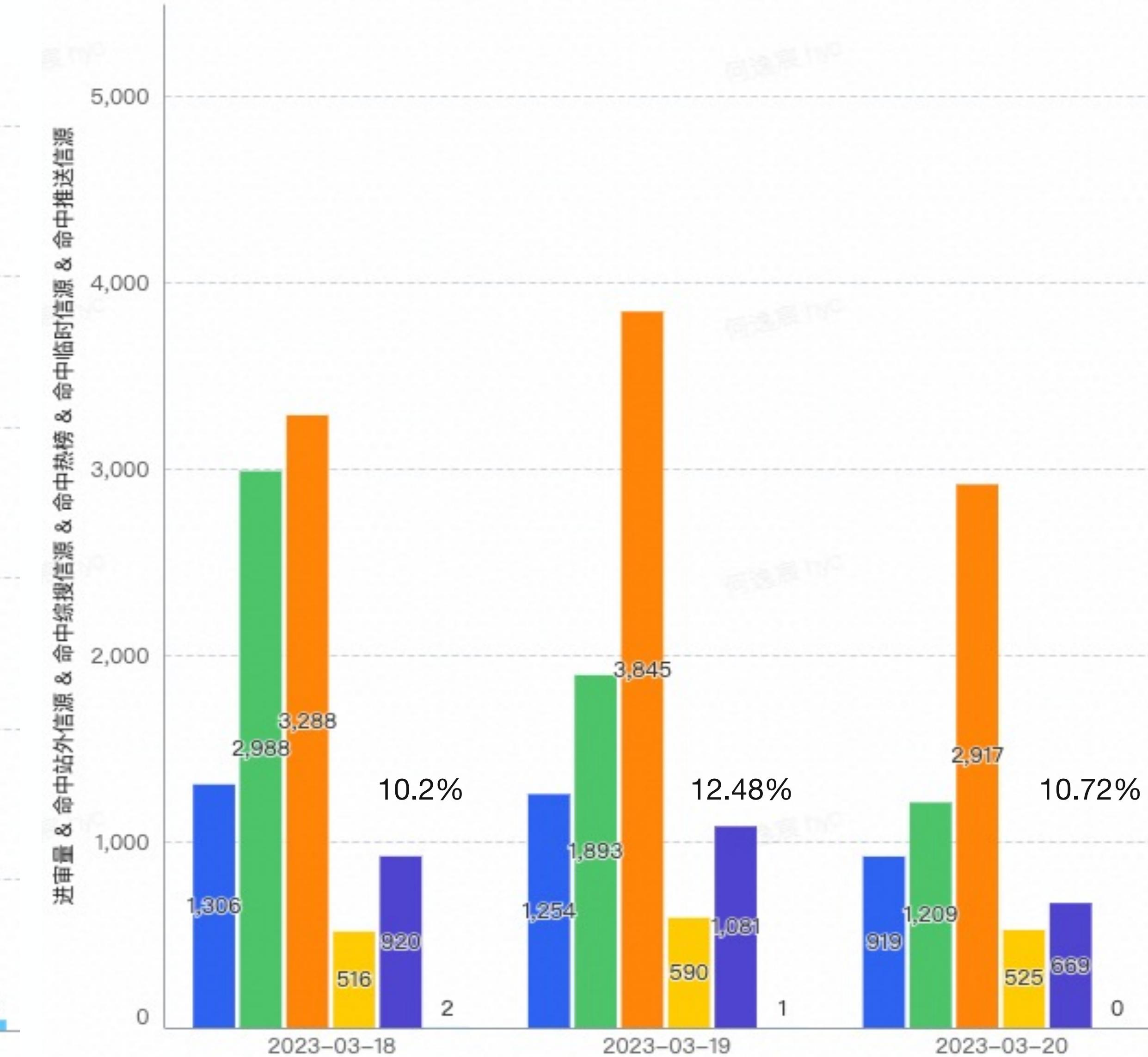
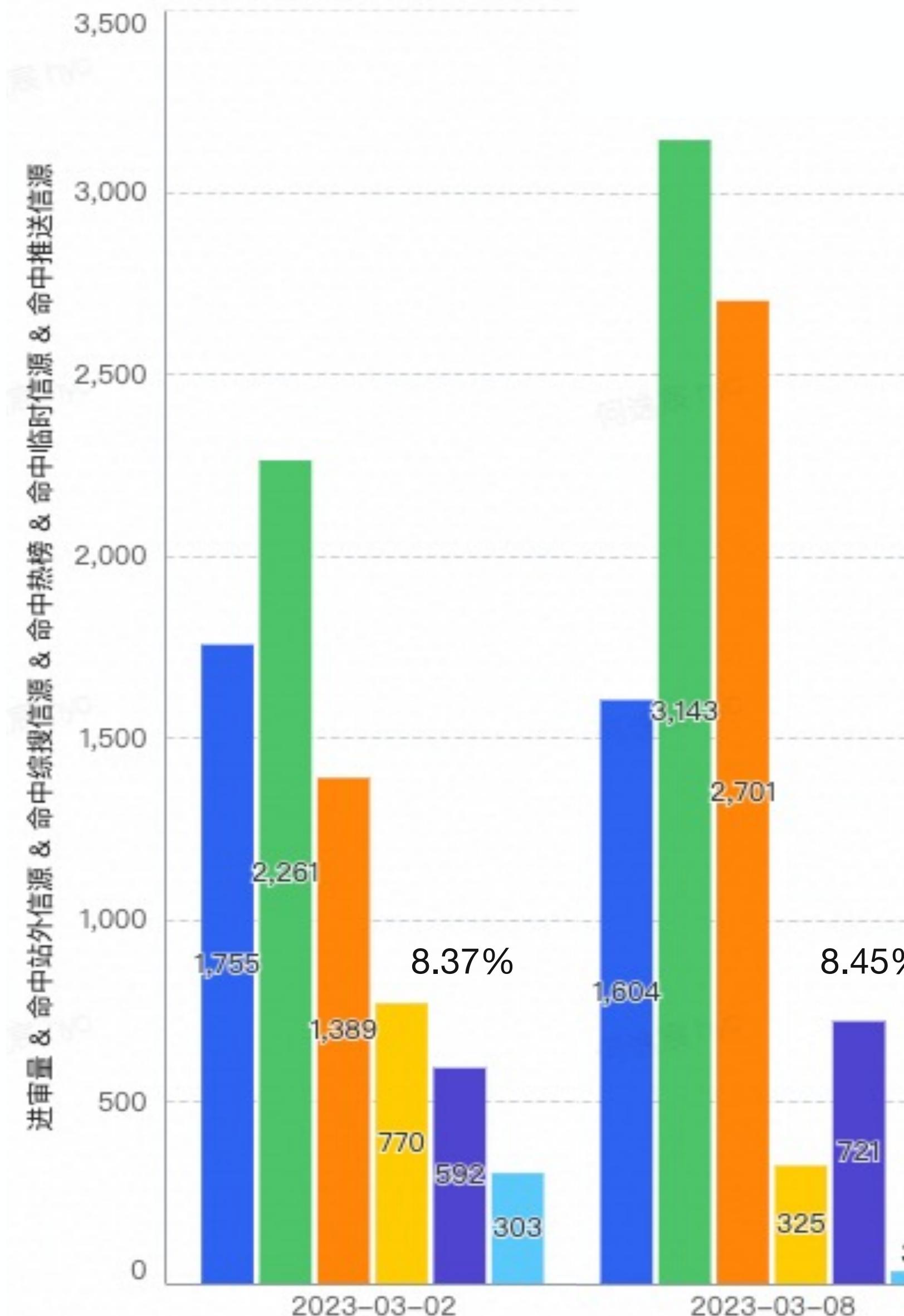


对于一些没有权威媒体报道的真实事件，可能会多次进审但不打压，降低队列打压率。

针对这一问题，我们构建了一个多次进审不打压的提示bot，每5分钟对于进审不打压的statement进行提示，运营同学根据需要录入临时信源库。

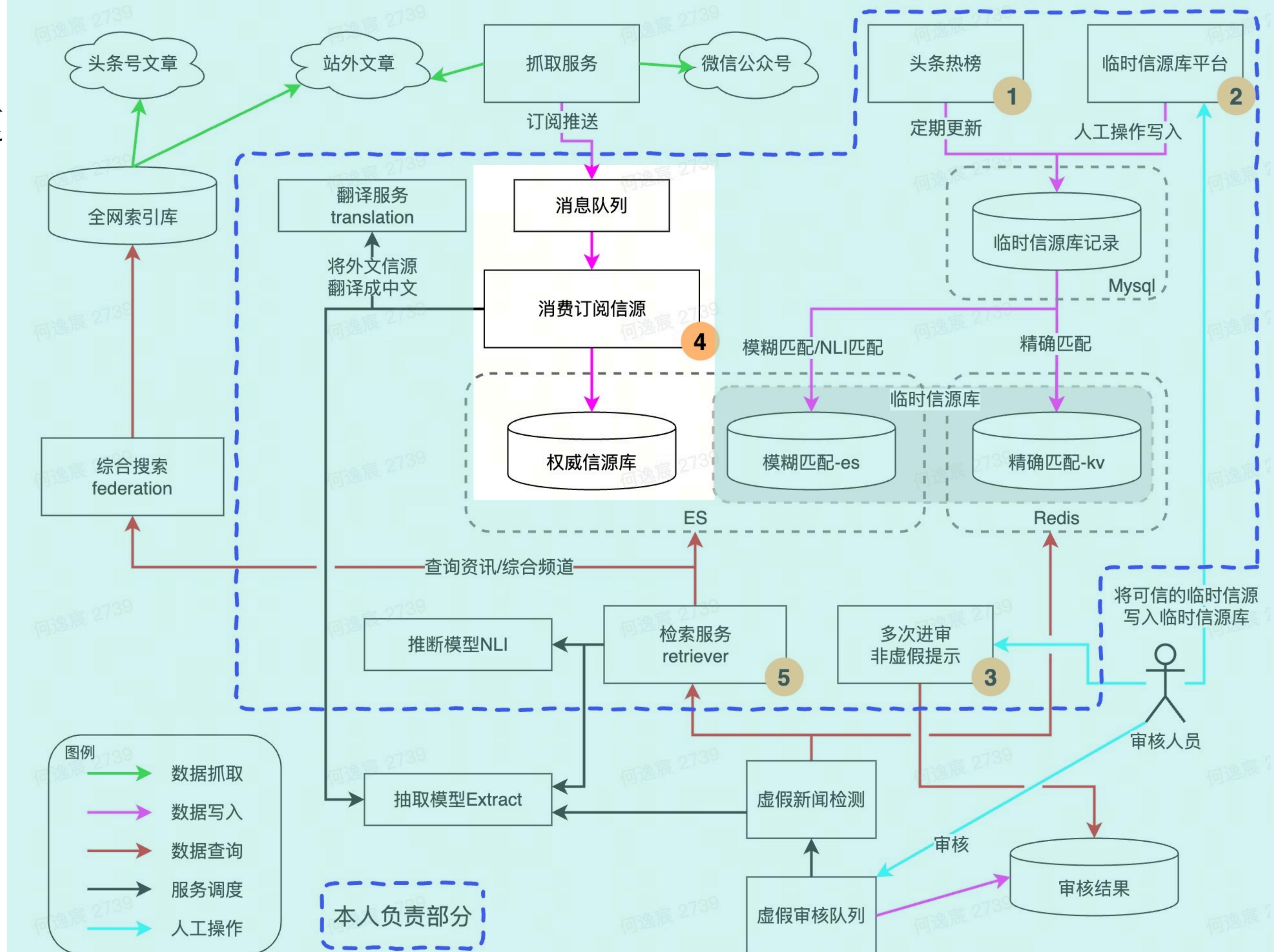
OK | 罗婧姝

# 多次进审非虚假提示 Python | MySQL | ES | CronJob

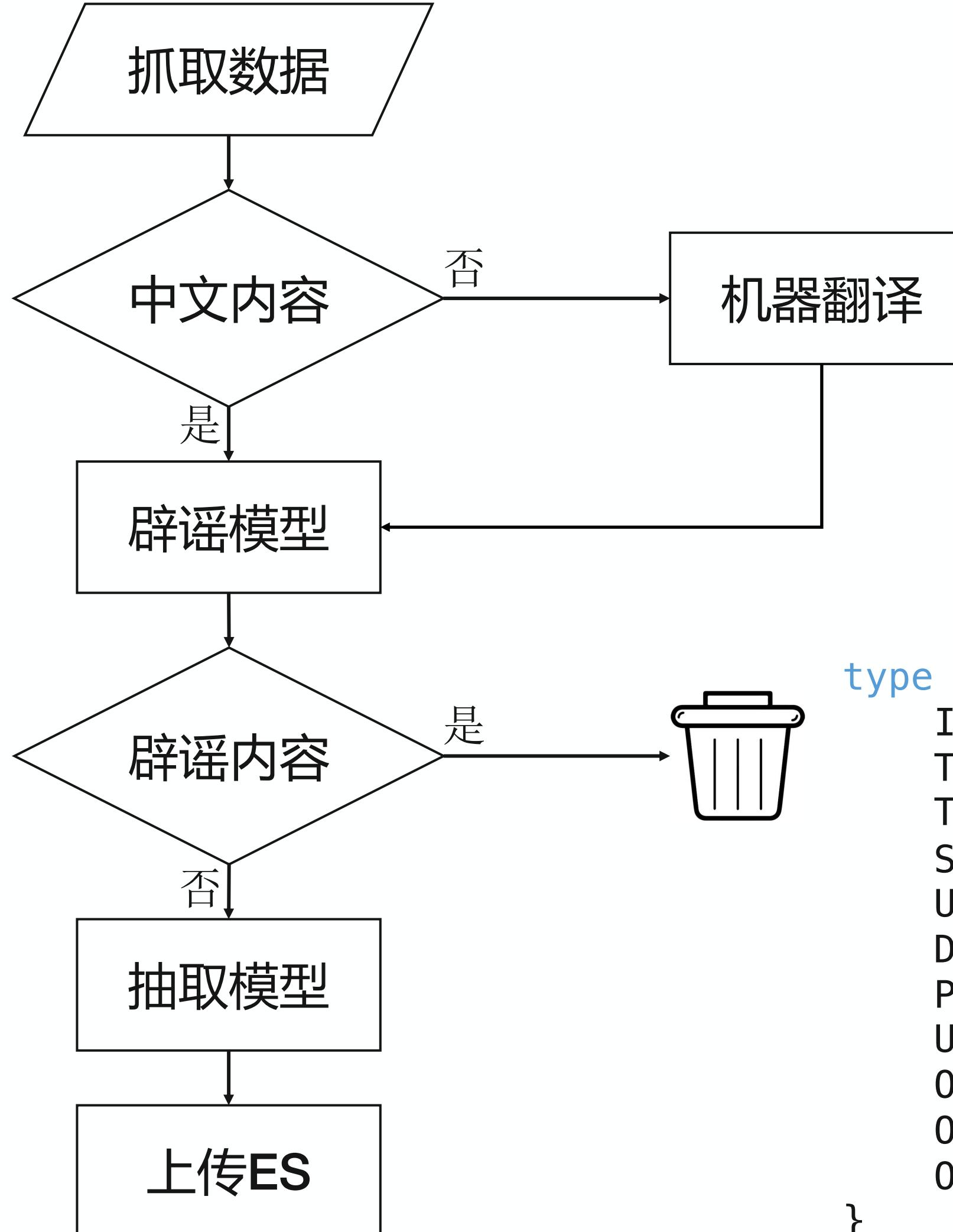


# 信源库系统搭建

1. 热榜内容自动入库
2. 临时信源库
3. 多次进审非虚假提示
4. 订阅信源消费
5. 检索服务



# 订阅信源消费 Golang | ES



订阅信源消费模块用于消费舆情中台抓取并写入kafka队列的信源，先前消费模块存在以下问题：

1. 使用Python语言编写，无法高并发消费，有消费滞后
  2. 对于辟谣内容，由于抽取不全导致谣言被当作信源
- 重构消费模块：

1. 使用Golang语言编写，支持并发消费，队列无堆积
2. 接入辟谣模型，识别到辟谣文章不录入信源库

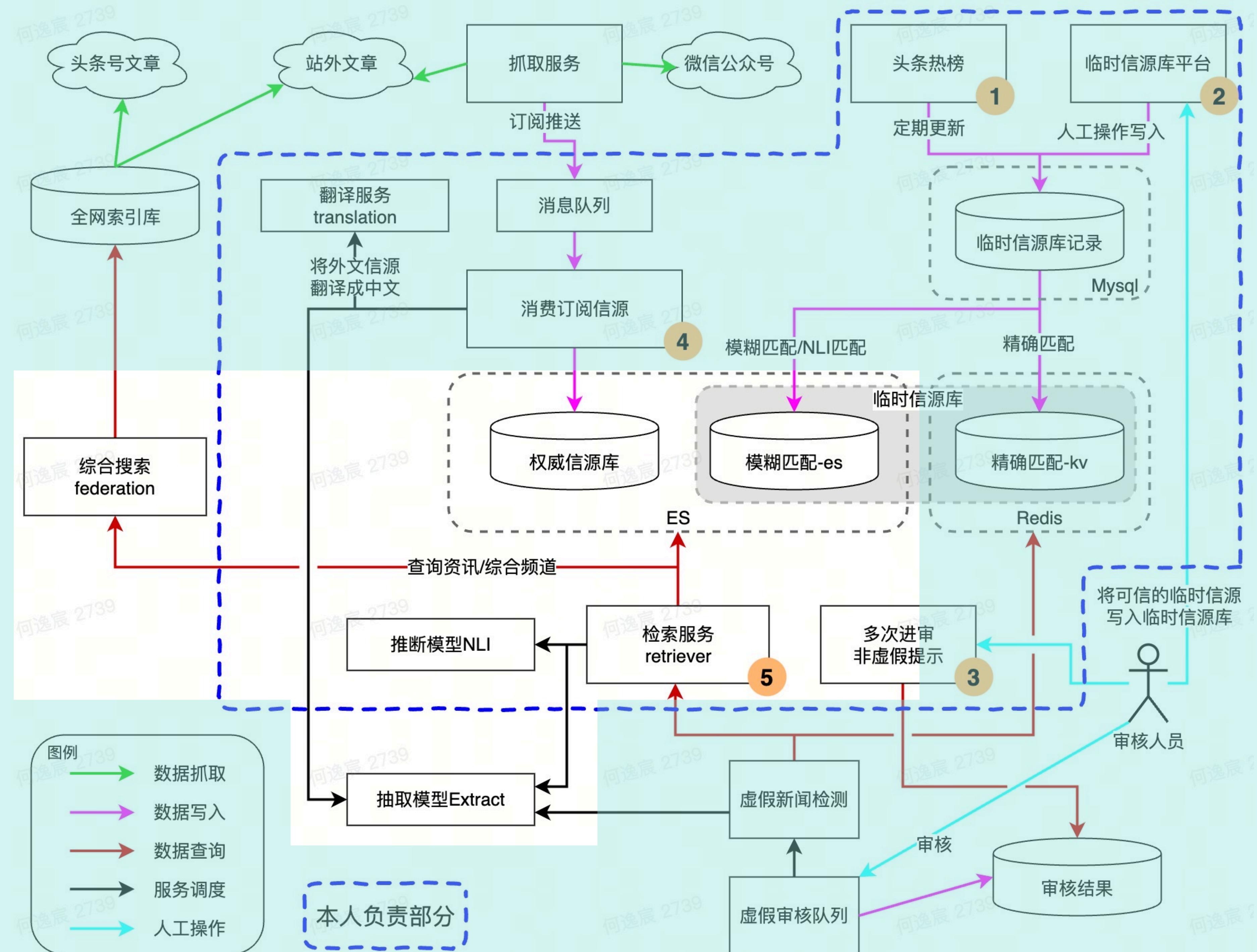
消费详情				
Partition ID	Latest Offset	Earliest Offset	lag	Group offset
0	2305803	2296315	0	2305799
1	2304876	2295011	0	2304876
2	2307851	2287276	0	2307849
3	2304542	2286630	0	2304541
4	2306233	2287978	0	2306229

```
type ESDoc struct {
    ID          uint64 `json:"id"`
    Title       string `json:"title"`
    Texts       []string `json:"texts"`
    Statements  []string `json:"statements"`
    URL         string `json:"url"`
    Domain     string `json:"domain"`
    PublishTime string `json:"publish_time"`
    UpdateTime  string `json:"update_time"`
    OriginalTitle string `json:"original_title"`
    OriginalContent string `json:"original_content"`
    OriginalTexts []string `json:"original_texts"`
}
```

The code defines the ESDoc struct for storing news documents. It includes fields for ID, Title, Texts, Statements, URL, Domain, PublishTime, UpdateTime, OriginalTitle, OriginalContent, and OriginalTexts. The struct uses JSON tags to map fields to their corresponding JSON keys.

# 信源库系统搭建

1. 热榜内容自动入库
2. 临时信源库
3. 多次进审非虚假提示
4. 订阅信源消费
5. 检索服务



# 检索服务 Golang

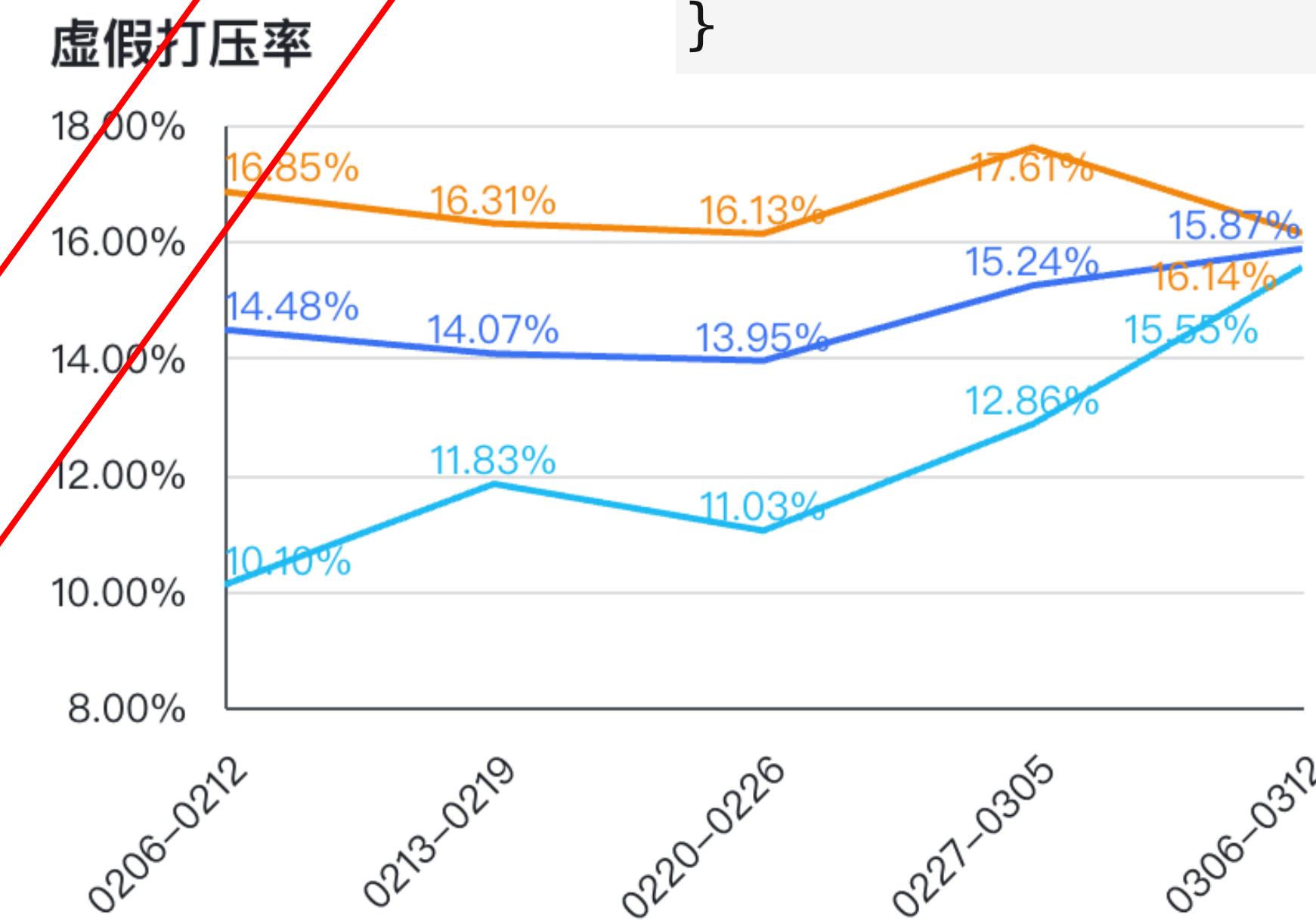
检索服务用于对进审的虚假新闻内容中的关键事件检索信源库，看是否有权威信源支撑。

先前服务存在以下问题：

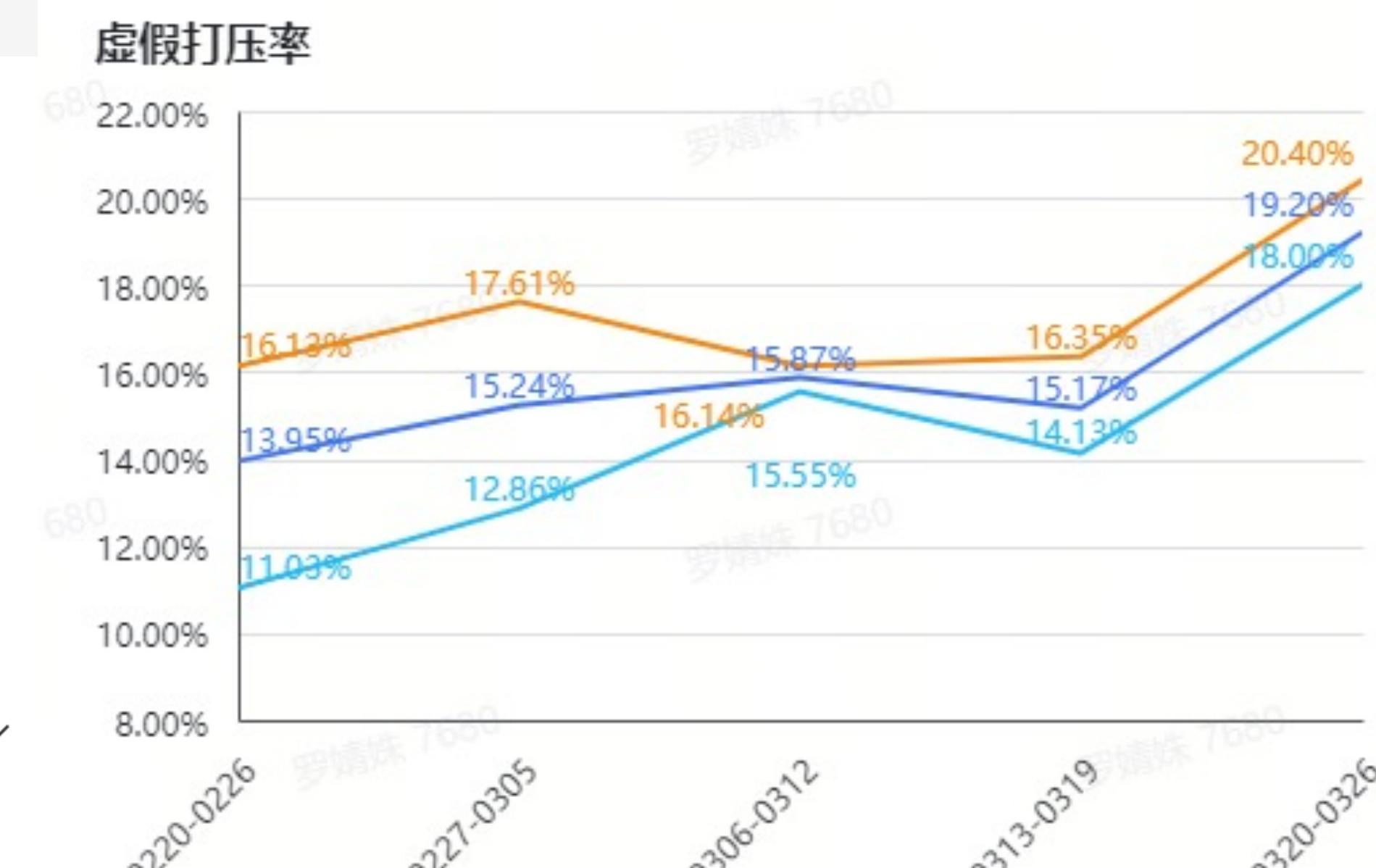
1. 只使用标题作为信源信息进行匹配
2. 仅使用es检索分数和qt分数判断是否相关

改进检索服务：

1. 对信源进行关键事件抽取，将抽取到的关键事件也作为权威信源。
2. 使用nli模型计算query和信源之间的语义相似度，并返回nli分数用于后续处理。



```
query: 12月3日, 国家宗教事务局令第17号公布了《互联网宗教信息服务管理办法》全文来了",  
"国家宗教事务局部署对宗教活动场所开展紧急排查", ...  
}  
,"statements": [  
    "《互联网宗教信息服务管理办法》全文来了",  
    "国家宗教事务局部署对宗教活动场所开展紧急排查", ...  
],  
,"nli_scores": [  
    {"pred": 2,  
    "prob": [  
        0.14509, 0.01017, 0.46734, 0.00007, 0.37733  
    ],  
    "statement": "《互联网宗教信息服务管理办法》全文来了"  
},  
    ...  
],  
...  
}
```



# 大模型训练

# 通用审核语言模型

## 业务场景大模型训练流程

### 当前风险内容管控难点

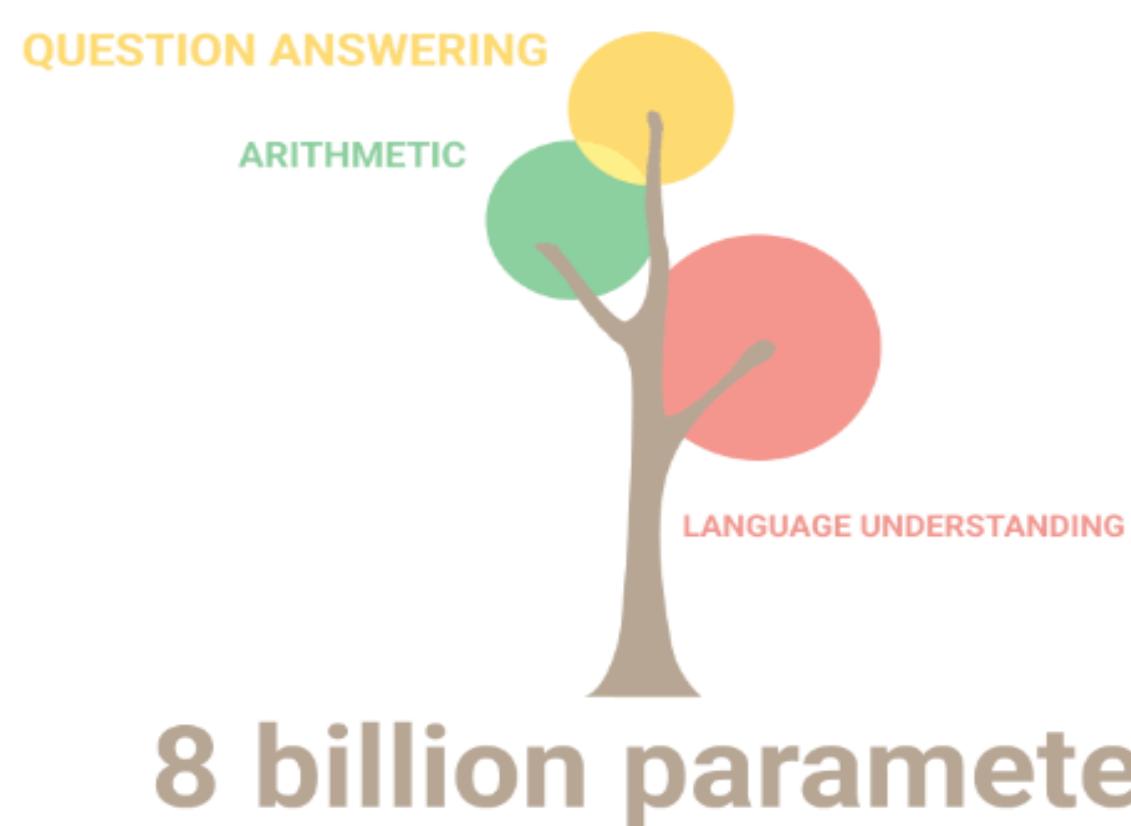
1. 紧扣热点事件，来不及训练专门的模型（“唐山打人事件”，“胡鑫宇事件”）
2. 管控内容较宽泛，需要深入理解语意（“炒作新冠造成大量死亡的言论”、“吹捧发达国家的言论”）

### 大模型使用的方法

B端私有部署活私域数据类服务，深度定制化应用

在小模型召回的基础上，增加大模型作为过滤，在保证尽量不漏招的情况下，提高模型的准确率。

在日常的迭代中，发现bad case后可以针对性构造几条数据添加到训练集中，模型即可有该方面的判断能力。



字节跳动

### 基底模型

**LLaMA:** Meta AI发布的一系列大规模语言模型( 7B / 13B / 33B / 65B )。尽管它在公开评测中取得了出色的绩，但中文的理解和生成能力都很差，且不具备通用的指令理解能力。

**LLaMA-plus:** AI Lab-NLP发起的一项LLaMA增强计划，通过多语言支持、通用指令理解、逻辑推理增强、知识增强、超长上下文、模型加速等方式对LLaMA。主要使用的版本包括：

- LLaMA-plus-0505 - 使用14B的平行语料和中文单语语料进行了预训练
- LLaMA-plus-0523 - 使用40B的平行语料和中文单语语料进行了预训练，在0505的基础上扩展 tokenizer (49139)，对中文的解码速度提升一倍

**gpt2lab:** AI Lab-NLP自研的GPT模型，基于原始的gpt-2进行改进，使用decoder-only的结构，词表大小为100032，使用250B tokens进行训练。

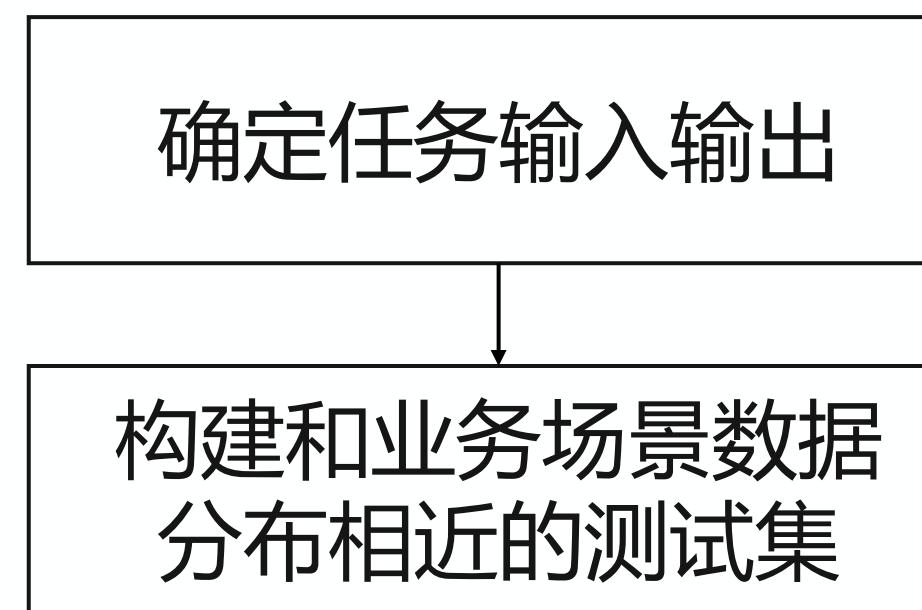
**vicuna:** 通过使用从ShareGPT.com获取的约70K个用户共享的会话对LLaMA基础模型进行微调。

**baichuan:** 百川智能发布的7B中英文预训练模型。训练数据以高质量中文语料为基础，同时融合了优质的英文数据，包含1200B tokens训练数据。

### 数据集

知识库	4.5k	在谣言识别场景下，请判断[文章]是否能推出[句子]，先解释原因，然后回答“是”或“否”。
指令库	6k	在指令识别场景下。请一步一步分析以下文本或视图内容与每个规则的相关性。
写作	98k	一些通用日常指令。
指令nli	264k	针对指令任务，请判断句子1和句子2的语义关系。
知识库nli	243k	针对谣言识别任务，请判断句子1和句子2的语义关系。
新闻nli	132k	针对虚假新闻识别任务，请判断句子1和句子2的语义关系。
风险识别	5k	请分析下列文本中的风险内容，列出风险内容以及对应的风险点，并提取重要事件。

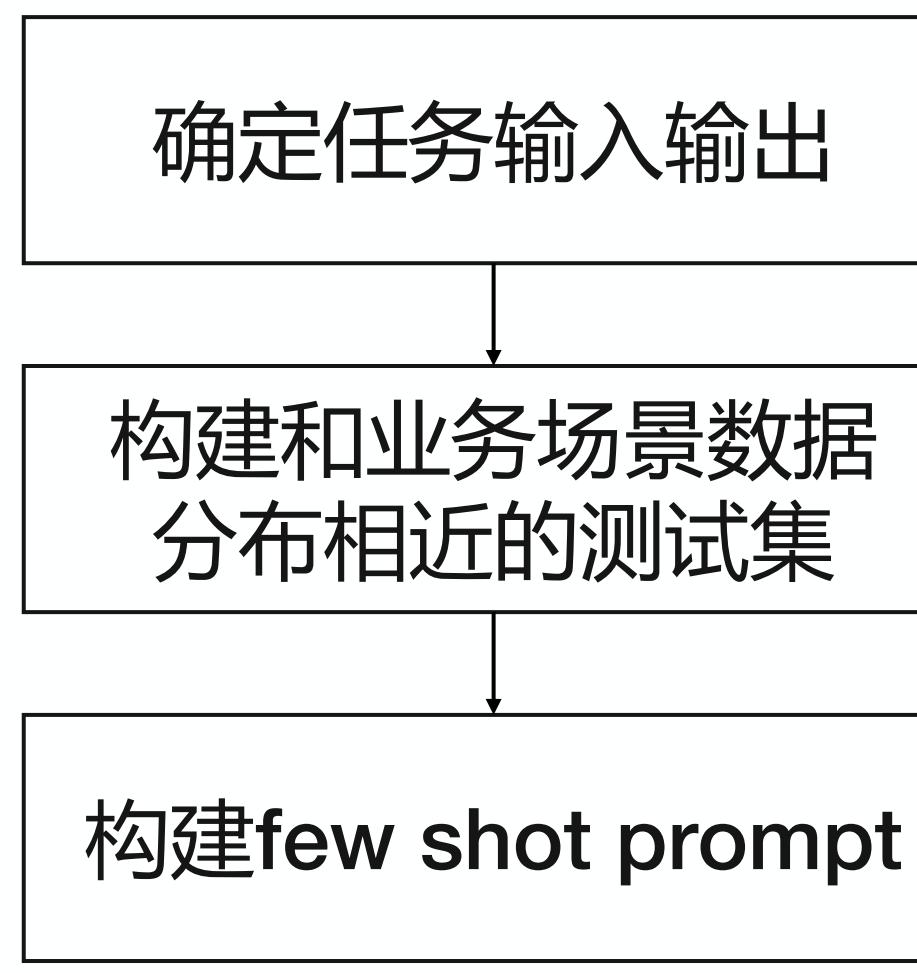
# 通用审核语言模型 训练数据构造 – 以知识库为例



优先构建测试集，为尝试**prompt**写法和模型训练迭代提高指标支持。和业务场景数据分布近似：

- 正负例比例近似1:1
- 正例按线上召回率采样进且打压**case**和漏招**case**
- 负例包括进审且不打压**case**和未进审**case** (3:2)

# 通用审核语言模型 训练数据构造 – 以知识库为例



优先构建测试集，为尝试**prompt**写法和模型训练迭代提高指标支持。和业务场景数据分布近似：

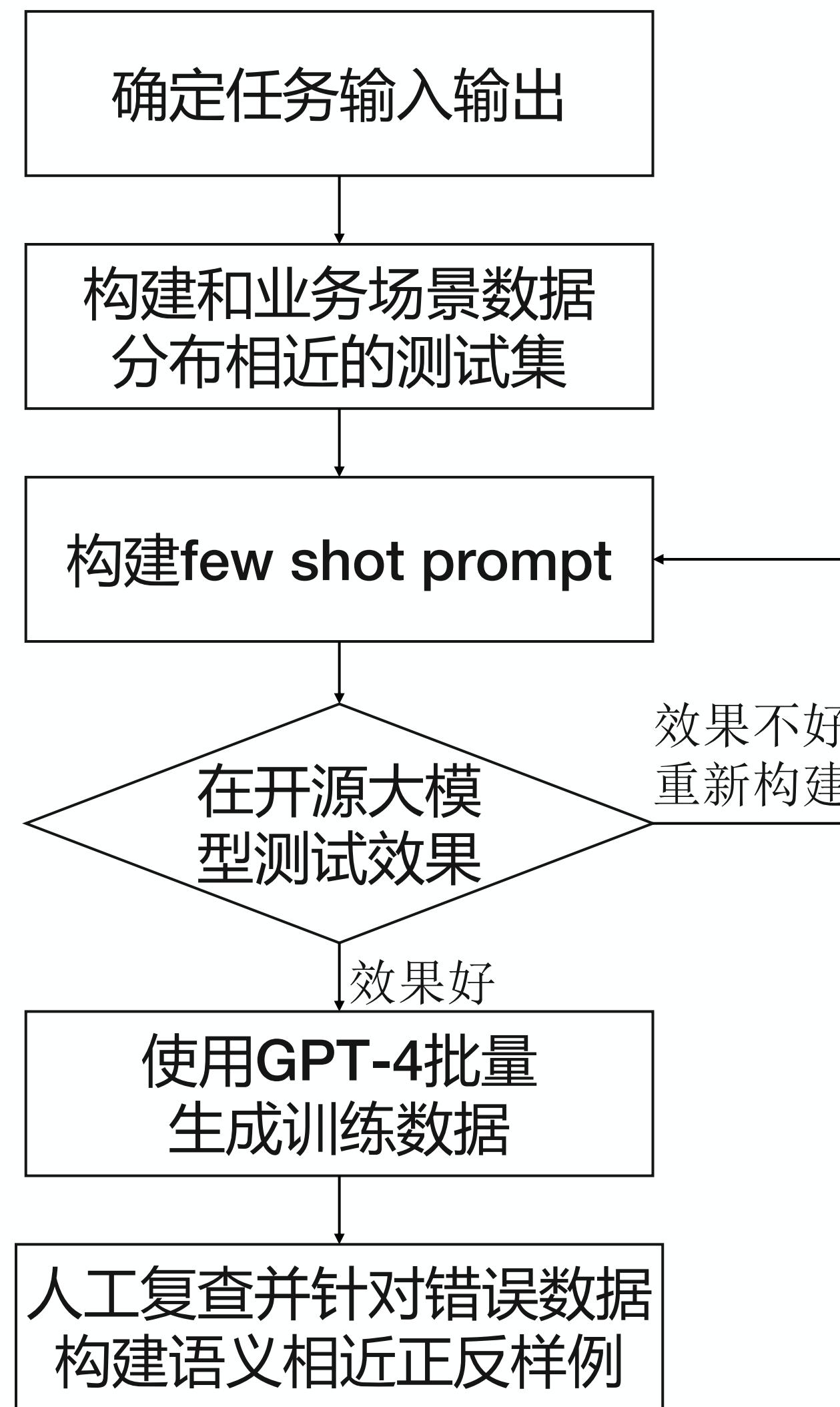
- 正负例比例近似1:1
- 正例按线上召回率采样进且打压**case**和漏招**case**
- 负例包括进审且不打压**case**和未进审**case** (3:2)

请判断[文章]是否能推出[句子]，先解释原因，然后回答“是”或“否”。

[文章] xxx

[句子] xxx

# 通用审核语言模型 训练数据构造 – 以知识库为例



优先构建测试集，为尝试prompt写法和模型训练迭代提高指标支持。和业务场景数据分布近似：

- 正负例比例近似1:1
- 正例按线上召回率采样进且打压case和漏招case
- 负例包括进审且不打压case和未进审case (3:2)

请判断[文章]是否能推出[句子]，先解释原因，然后回答“是”或“否”。

[文章] xxx

[句子] xxx

[文章] 生姜还有杀灭口腔致病菌和肠道致病菌的作用，用生姜水含漱治疗口臭和牙周炎，有一定的疗效。大院姜农余丰收正在收获成熟的生姜 六、白姜的吃法和注意事项  
铜陵白姜吃法很多，有喝姜汤（姜汁），吃姜粥等，糖冰姜、姜茶、姜炒制菜肴的食用更是司空见惯。既能使味道鲜美，又有助于开胃健脾，促进食欲，帮助消化。  
[句子] 生姜水漱口治疗牙周炎

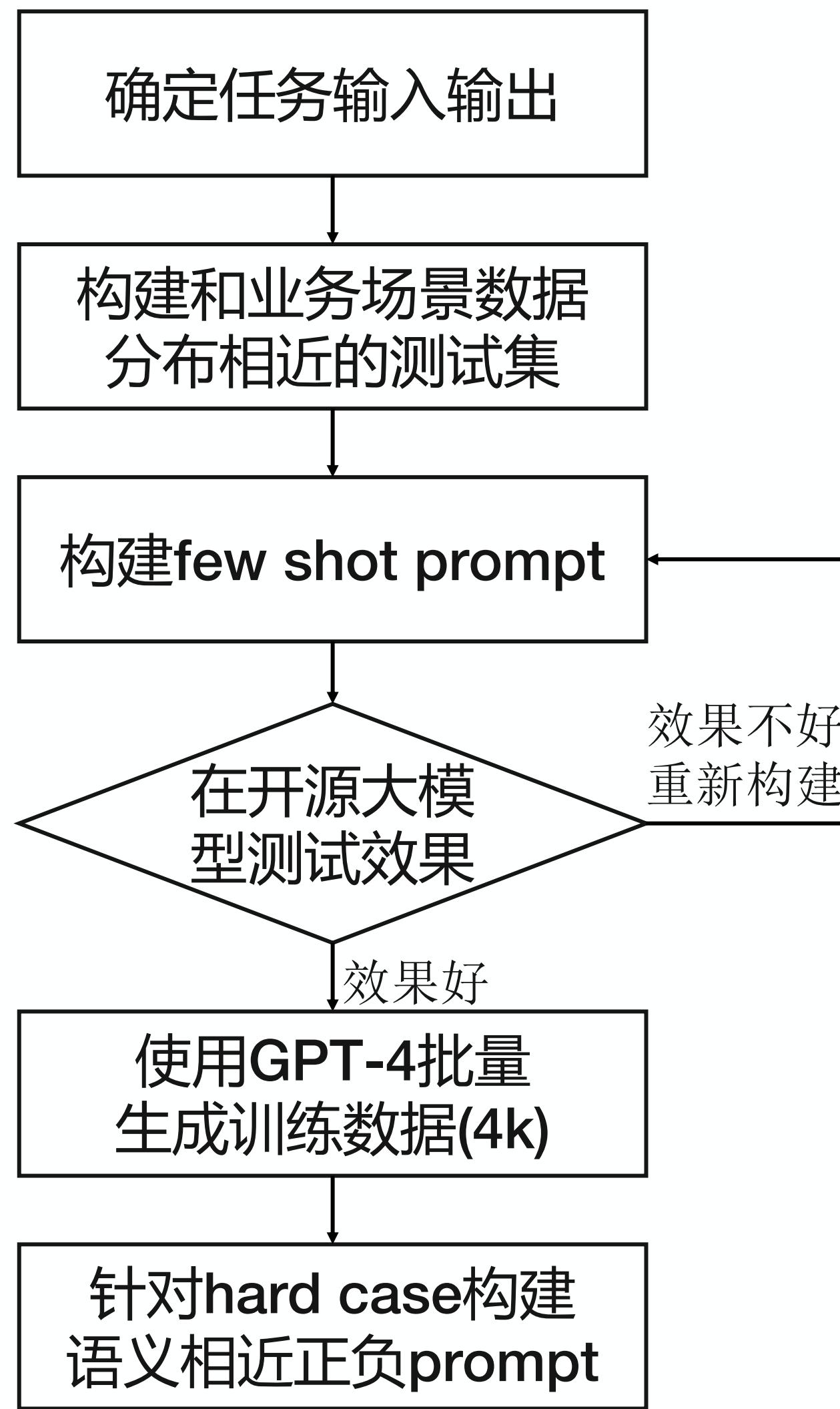
判断依据：

1. [文章]提到用生姜水含漱治疗口臭和牙周炎，有一定的疗效。
2. [句子]提到生姜水漱口治疗牙周炎。
3. 用生姜水含漱治疗口臭和牙周炎可以视为生姜水漱口治疗牙周炎的具体应用。
4. [文章]可以推出[句子]，是。

# 通用审核语言模型

## 训练数据构造 - 以知识库为例

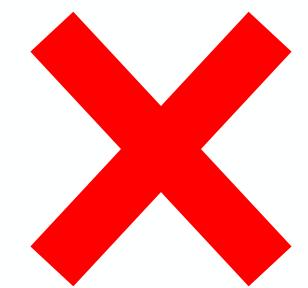
帮助模型更好地理解业务判断尺度  
提升解决疑难问题的能力



【文章】生育无限制—马来西亚不仅没有计划生育规定，而且政府鼓励多生、优生。想生就生，不受限制。拿到红卡后新生儿将自动获得大马国籍。

【谣言点】在马来西亚出生的孩子就可拥有马来西亚护照或国籍  
判断依据：

1. 【文章】提到拿到马来西亚红卡后新生儿将自动获得大马国籍。
2. 【句子】提到在马来西亚出生的孩子就可拥有马来西亚护照或国籍。
3. 拥有马来西亚国籍意味着可以拥有马来西亚护照。
4. 【文章】可以推出【句子】，是。



【文章】生育无限制—马来西亚不仅没有计划生育规定，而且政府鼓励多生、优生。想生就生，不受限制。拿到红卡后新生儿将自动获得大马国籍。

【谣言点】在马来西亚出生的孩子就可拥有马来西亚护照或国籍  
判断依据：

1. 【文章】提到拿到马来西亚红卡后新生儿将自动获得大马国籍。
2. 【句子】提到在马来西亚出生的孩子就可拥有马来西亚护照或国籍。
3. 不是所有新生儿都能拿到马来西亚国籍，只有父母拿到红卡后，新生儿才能拥有马来西亚国籍。
4. 【文章】不能推出【句子】，否。

【文章】生育无限制—马来西亚不仅没有计划生育规定，而且政府鼓励多生、优生。想生就生，不受限制。所有新生儿将自动获得大马国籍。

【谣言点】在马来西亚出生的孩子就可拥有马来西亚护照或国籍  
判断依据：

1. 【文章】提到拿到马来西亚红卡后新生儿将自动获得大马国籍。
2. 【句子】提到在马来西亚出生的孩子就可拥有马来西亚护照或国籍。
3. 拥有马来西亚国籍意味着可以拥有马来西亚护照。
4. 【文章】可以推出【句子】，是。

# 通用审核语言模型 训练数据构造 - 以知识库为例

## 实验效果

model	precision	recall	F1 score	备注
online albert	0.558673469	0.958424508	0.705882353	
GPT4 few shot	0.699490662	0.901531729	0.787762906	
llama13b	0.697628458	0.772428884	0.733125649	知识库训练数据 翻译成英文
llama30b	0.7208413	0.824945295	0.769387755	原始llama，只能接受英文输入，将中文通过现有翻译后输入，翻译过程有损失。
llama65b	0.723684211	0.842450766	0.778564206	
llama-plus 13b 0505	0.725646123	0.79868709	0.760416667	中文预训练模型
gpt2lab 13b	0.724899598	0.789934354	0.756020942	知识库训练数据
llama-plus 13b 0523	0.720149254	0.84463895	0.777442095	知识库+指令库+0.1写作数据 训练数据
gpt2lab 13b	0.718867925	0.833698031	0.772036474	将知识库和指令库训练数据的prompt进行多样化改写，提升模型对prompt的理解能力
llama-plus 65b	0.771929825	0.866520788	0.816494845	
llama-plus 13b 0523	0.751984127	0.829321663	0.788761707	知识库+指令库+风险事件 训练数据

# 其他工作

实体链接  
健康专审

# 实体链接

背景：

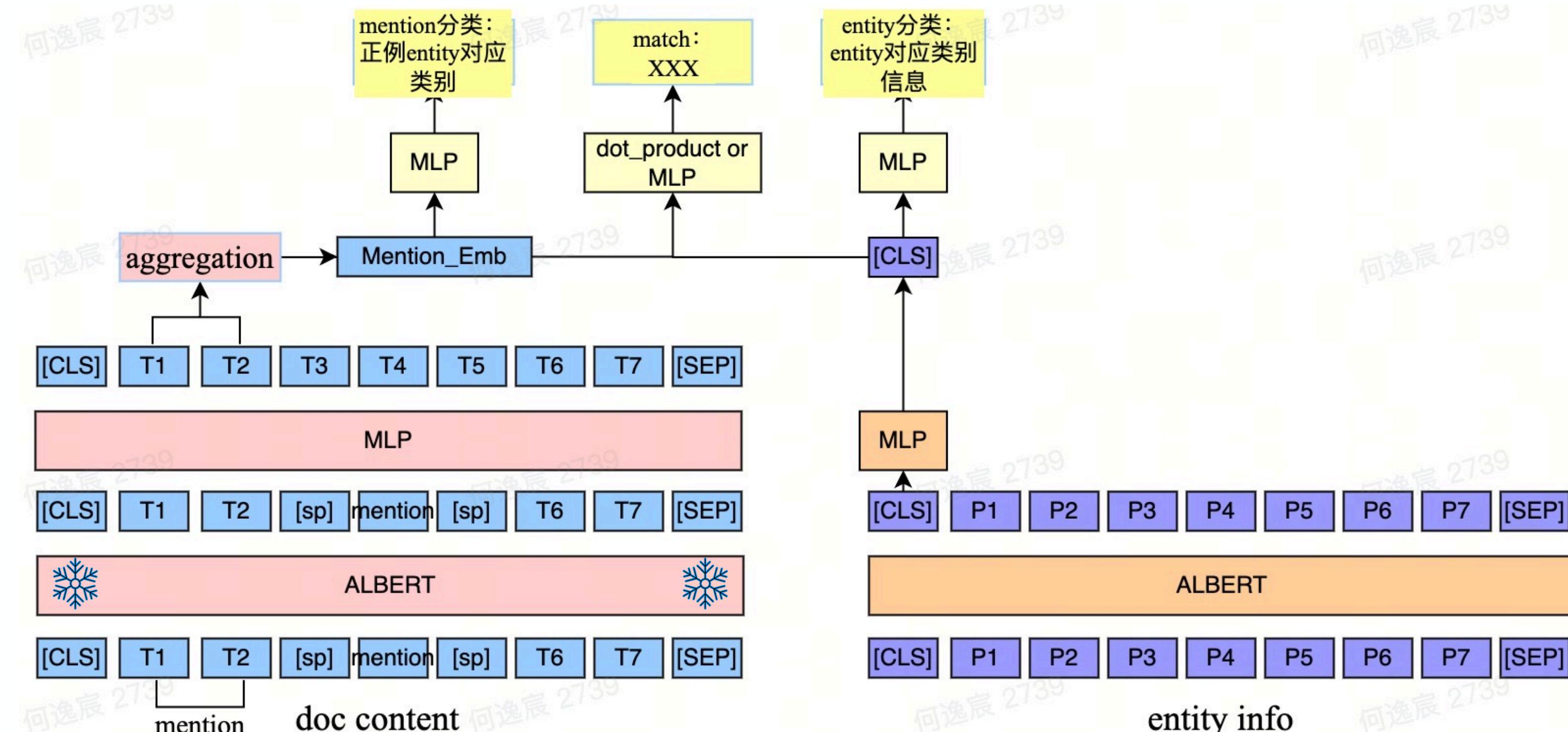
在尽量不影响精准问答在线模型效果的同时，将已有的实体链接模型的能力迁移到精准问答在线模型上。

方法：

- ☒ 1. 实体链接采用双塔模型，其中一个塔选择精准问答模型的encoder并freeze，训练entity info的encoder和上层的MLP。效果不佳，因为freeze的模型已经经过大量finetune，模型参数已拟合下游任务，原始语义信息较少。
- ✓ 2. 训练一个规模较大的teacher模型，并和精准问答其他任务一起进行蒸馏。

降维：

entity info侧的所有实体的embedding需要输出成tensor存在显存中，然而768维的tensor存不下，因此需要将768维的向量降维成128维，**使用PCA初始化降维矩阵可以对模型效果有很大的提高**。



# 健康专审

问题和难点：

伪科学文章通常可以通过写作风格和用词特点等篇章级特征进行鉴别，然而健康类文章涉及到许多细节用词和医学专业理论问题，很难通过整体文章输入进行二分类

基于资质的二分类模型：

从健康专审队列标准来看很多内容是完全和健康资质挂钩，因此将健康资质作为一个后处理部分，针对不同健康资质使用不同的阈值。使用资质2的数据作为训练数据，并在资质0和1上适当降低进审阈值。

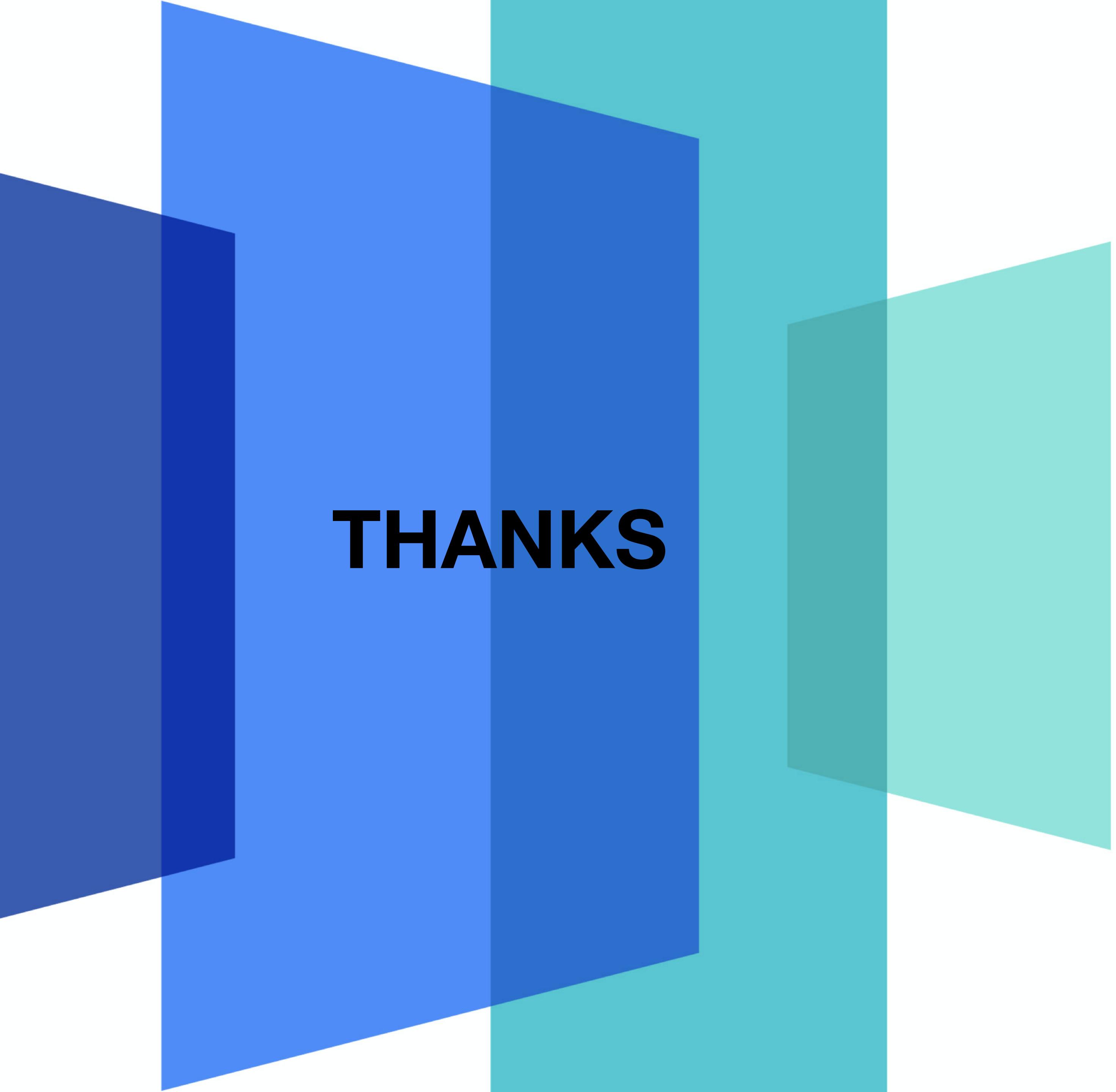
threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65
quality=0	0.47	<b>0.49</b>	0.484	0.473	0.454	0.425	0.403	0.362	0.334	0.301	0.255	0.225	0.188	0.15
quality=1	0.189	0.342	0.381	0.401	0.42	0.444	0.461	<b>0.465</b>	0.438	0.435	0.44	0.411	0.396	0.36
quality=2	0.341	0.464	0.512	0.522	0.533	0.546	0.551	0.548	0.533	0.531	<b>0.549</b>	0.53	0.51	0.495

0 - 没有健康资质  
1 - 普通健康资质  
2 - 专业健康资质

根据试验结果，分别对资质0和资质1降低阈值到0.05和0.35最佳

- 对于资质0，预测分数>0.05进审
- 对于资质1，预测分数>0.35进审
- 对于资质2，预测分数>0.5进审

该项目目前模型测ready，等待排期  
 ByteDance 字节跳动



THANKS

