

内容质量项目

字节跳动 - AI Lab
2022.5 - 至今

01

虚假信息检测

02

伪科学内容识别

03

信源库系统搭建

04

通用审核模型构建

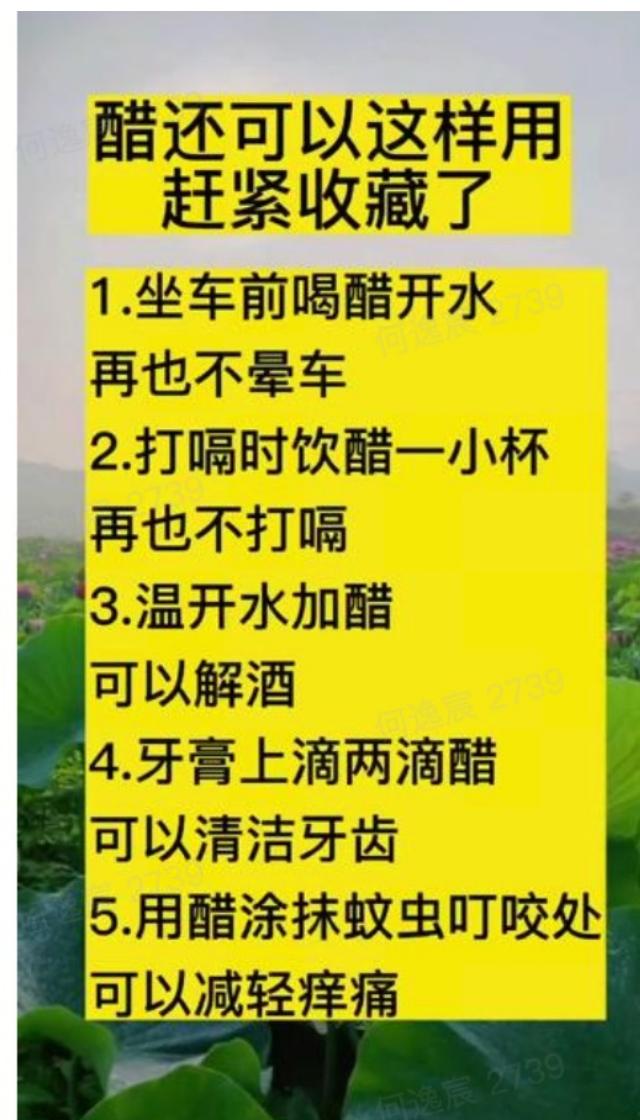
虚假信息检测 — 背景

- 影响对重大事件的认知 -- 虚假新闻

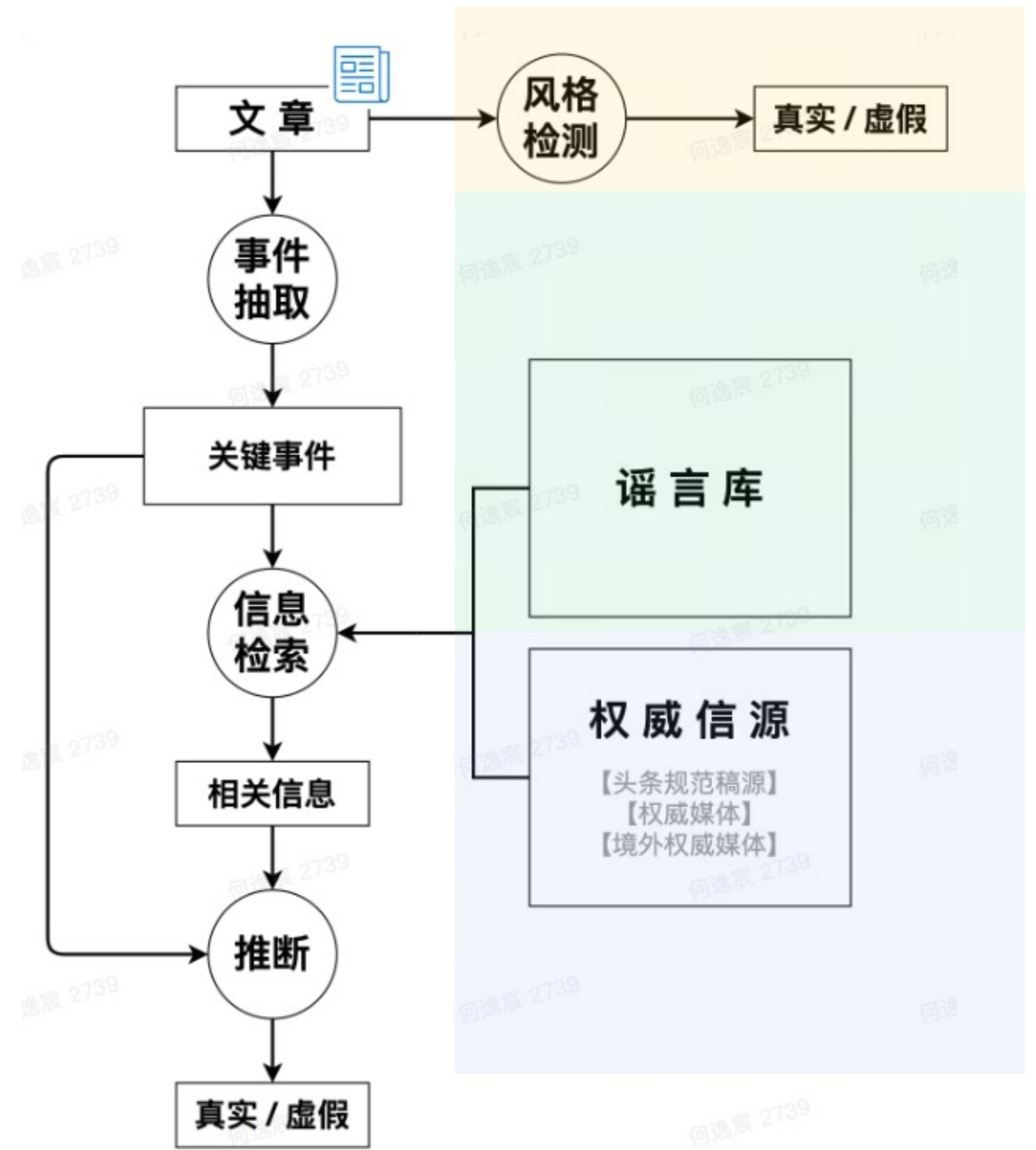
俄乌冲突

- 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
- 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

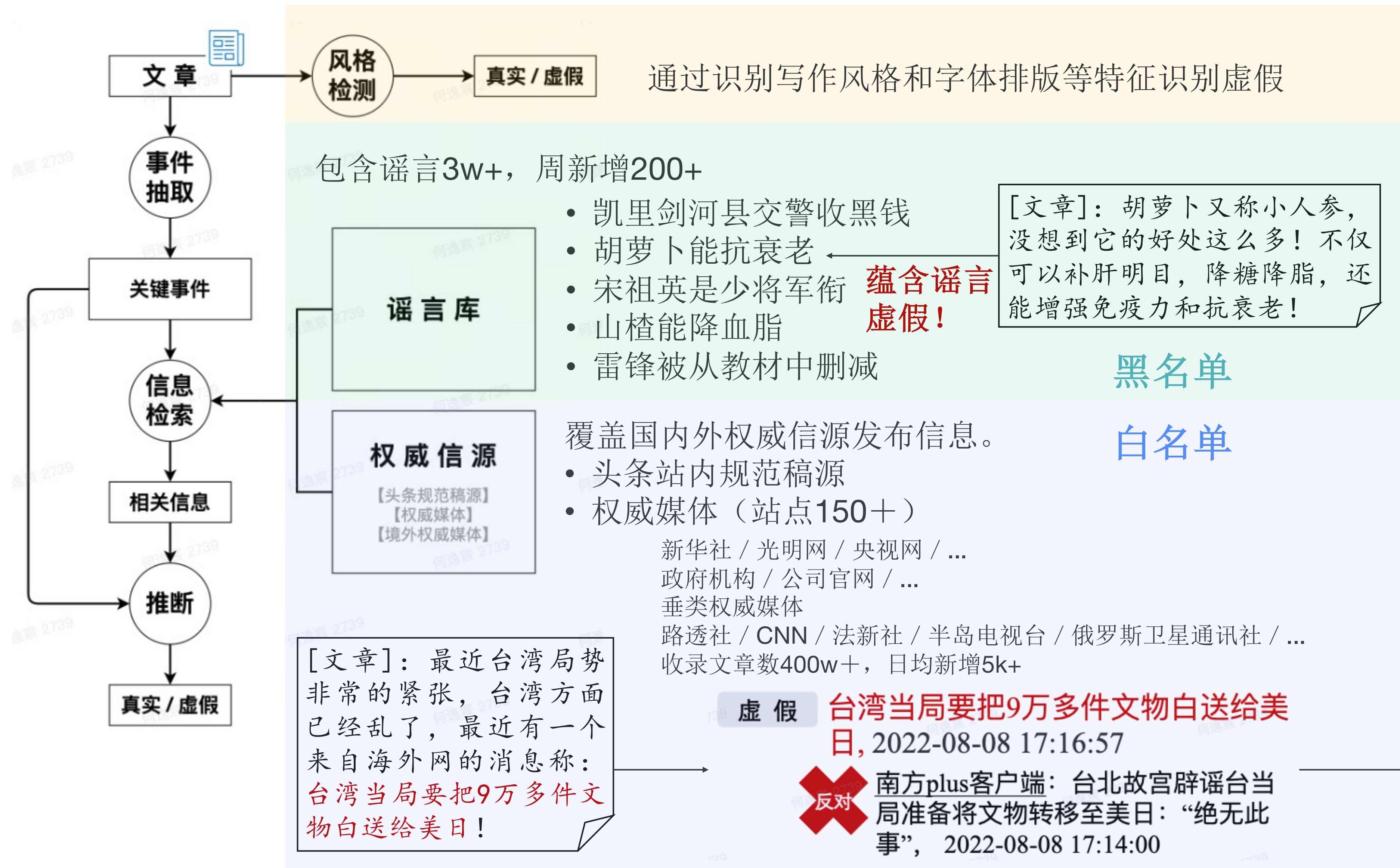
- 威胁健康，造成财产损失 -- 伪科学、谣言库



虚假信息检测 — 方法



虚假信息检测 — 方法



本人从事的工作：

伪科学内容识别

信源库系统搭建

通用审核模型

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

01

虚假信息检测

02

伪科学内容识别

03

信源库系统搭建

04

通用审核模型构建

伪科学内容识别 — 意义

伪科学谣言为什么会被传播

内容披着“科学”外衣，具有强迷惑性

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”

清清宇泉 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、超轻水。科学研究发现，饮用水的氘含量越低，水的活性越强，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 何逸辰 2739 评论 2 暂无评论，点击抢沙发

“院士只会讲脱口秀”

普星撞火星 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 何逸辰 2739 评论 2 暂无评论，点击抢沙发

“外星人到过地球”

外星人在5000年前就出现过？史前壁画记载外星人早已拜访地球

岛主阿坤 2021-11-16 12:13

这一切神秘事件有些与牛鬼蛇神有关，有些与物理现象有关，有些与外星人有关。今天我们要讲的，是与外星人有牵连的事情。大多数情况下，将某件事视为外星人访问地球只是人类的一种幻想，而目前人们所能做的只有确认这神秘本身的真实性。

何逸辰 2739 评论 2 暂无评论，点击抢沙发

“高速物质形成金属氢”

金童希瑞 2021-12-27 15:33

小行星俯冲瞬间高速流动的物质转化为金属氢，金属氢聚合形成的二氧化硅以及二氧化硅衍生的硅酸盐会记录小行星俯冲瞬间形成的磁场。

何逸辰 2739 评论 2 暂无评论，点击抢沙发

“地球在不停地摇摆”

摇摆的地球 2021-12-27 06:22

地轴倾斜角是地球摇摆运动形成的，倾斜角度即是地球的摇摆幅度，地轴以约183天为一个周期缓慢旋转，致使地球出现四季变化。伴随地球公转，地球要完成两个摇摆运动周期，地球表面才能出现一次四季变化，如果没有公转，只有自转和摇摆运动，只需完成一个摇摆运动周期，地球表面就能出现一次四季变化

四季变化是地球摇摆运动形成的，地轴倾斜方向是不停的匀速的旋转改变的，所谓的“地球公转形成四季变化”理论和“地轴总是倾斜的指向同一方向”理论都是绝对错误的。

何逸辰 2739 评论 2 暂无评论，点击抢沙发

伪科学内容识别 — 方法

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，甚至是不可见的，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
郭平	校长	郑州市春霖职业培训学校
白大勇	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
李松林	原国家地震局兰州物探中心主任	地震监测站
白玉忠	主任医师	河南医学院高等专科学校附属医院
尹杰		白衣天使志愿者
高爱华	家长	春霖观察者(志愿者)
孙静霞	家长	孙嘉宝妈妈
赵真真	家长	陈静墨妈妈
和大勇	教师	新郑市道德模范小学员



图2.

鸡蛋矿物质组成，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^[2]。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^[2]。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^[2-3]。

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。

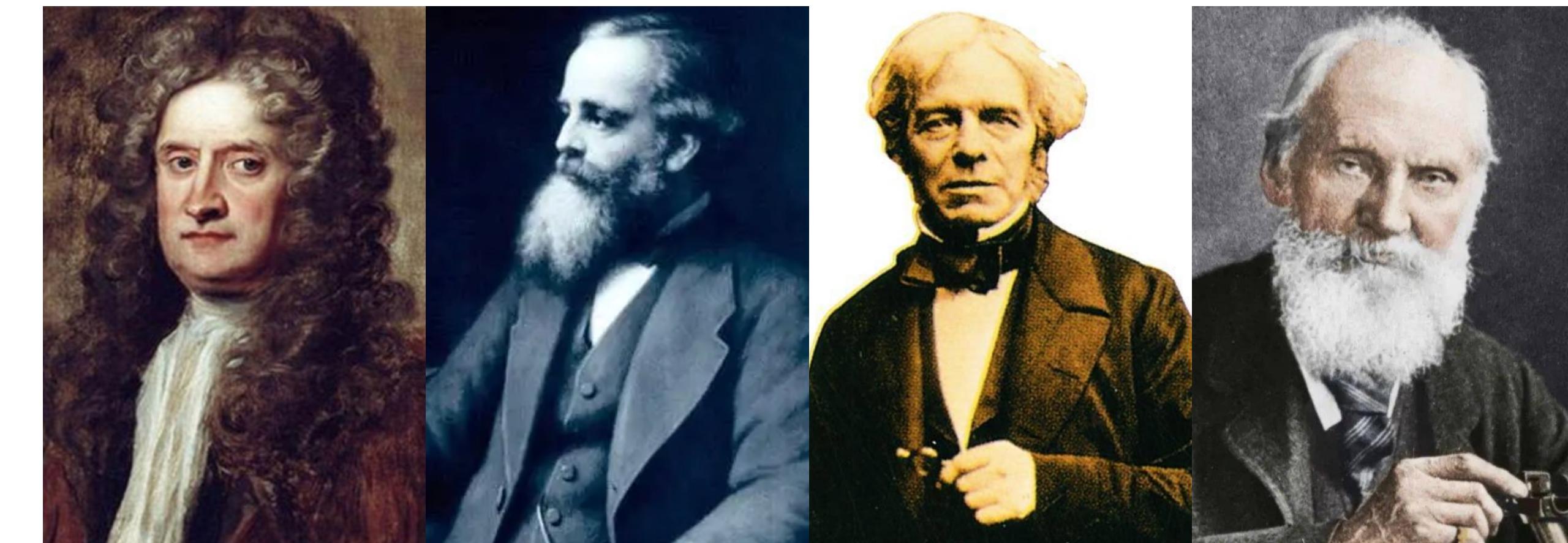
事实上，很多大名鼎鼎的“老”英国科学家都信神，比如近代物理学之父牛顿、统计物理学的奠基人之一麦克斯韦、电学之父和交流电之父法拉第、热力学之父开尔文勋爵……

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。

其他人都只是拉来活跃气氛的，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。



伪科学内容识别 — 方法

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

摘要
熟鸡蛋在郑州大学实验室中孵出小鸡，从而得出结论。

关键词
【中图分】

鸡蛋奇

蛋通过高温

用自己的超

蛋，现在我们

探究其内在

实验材

实验方

实验时

室内温

参加人

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。**

><加粗>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。 </加粗><插图>

姓名	职称	部门
马建民	院长	郑州大学设计学院副院长
李松海	原国家地震局郑州地质中心主任	地震监测中心
白卫云	主任医师	河南中医药高等专科学附属医院
尹杰	院长	山东中医药大学
肖爱华	院长	唐氏基因有限公司
孙静霞	院长	孙静霞诊所
赵真真	院长	碧彩墨绣坊
郭平	教师	新郑市实验小学



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。
① 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。
② 胚盘：位于卵黄表面中央有一圆形盘状小白点，呈稍凹状等。
③ 卵白：卵壳膜与卵黄膜之间，为胚胎发育提供水和营养物质。
④ 卵壳膜：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，胚盘色浅而小，已受精的卵，胚盘色浓而略大，这是因为胚胎发育已经开始^[1-2]，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^[1-2]。

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23分钟，整体上鸡蛋内部均有液态变成固态，在返生过程中，不添加任何化学物质，不进行任何物理处理，如加热或者降温，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，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**。

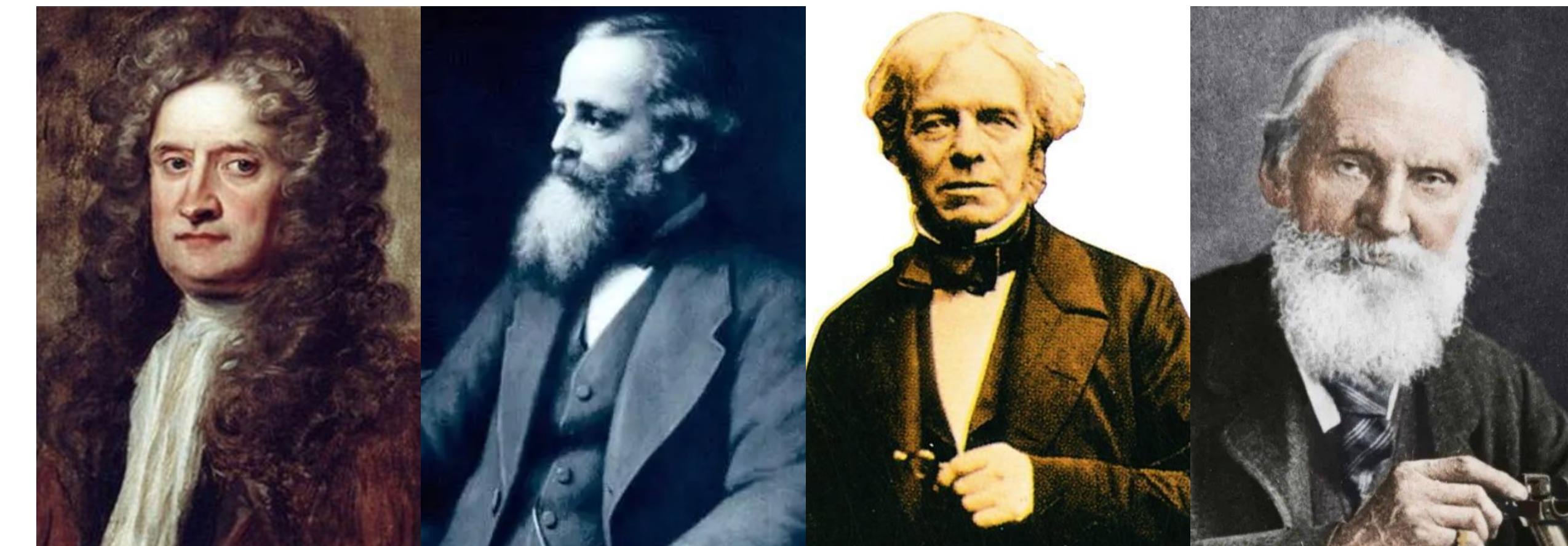
事实上，很多大名鼎鼎的“老”英国科学家都信神，比如近代物理学之父牛顿、统计物理学的奠基人之一麦克斯韦、电学之父和交流电之父法拉第、热力学之父开尔文勋爵……

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。

其他人都只是拉来活跃气氛的，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。



伪科学内容识别 — 方法

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

训练流程

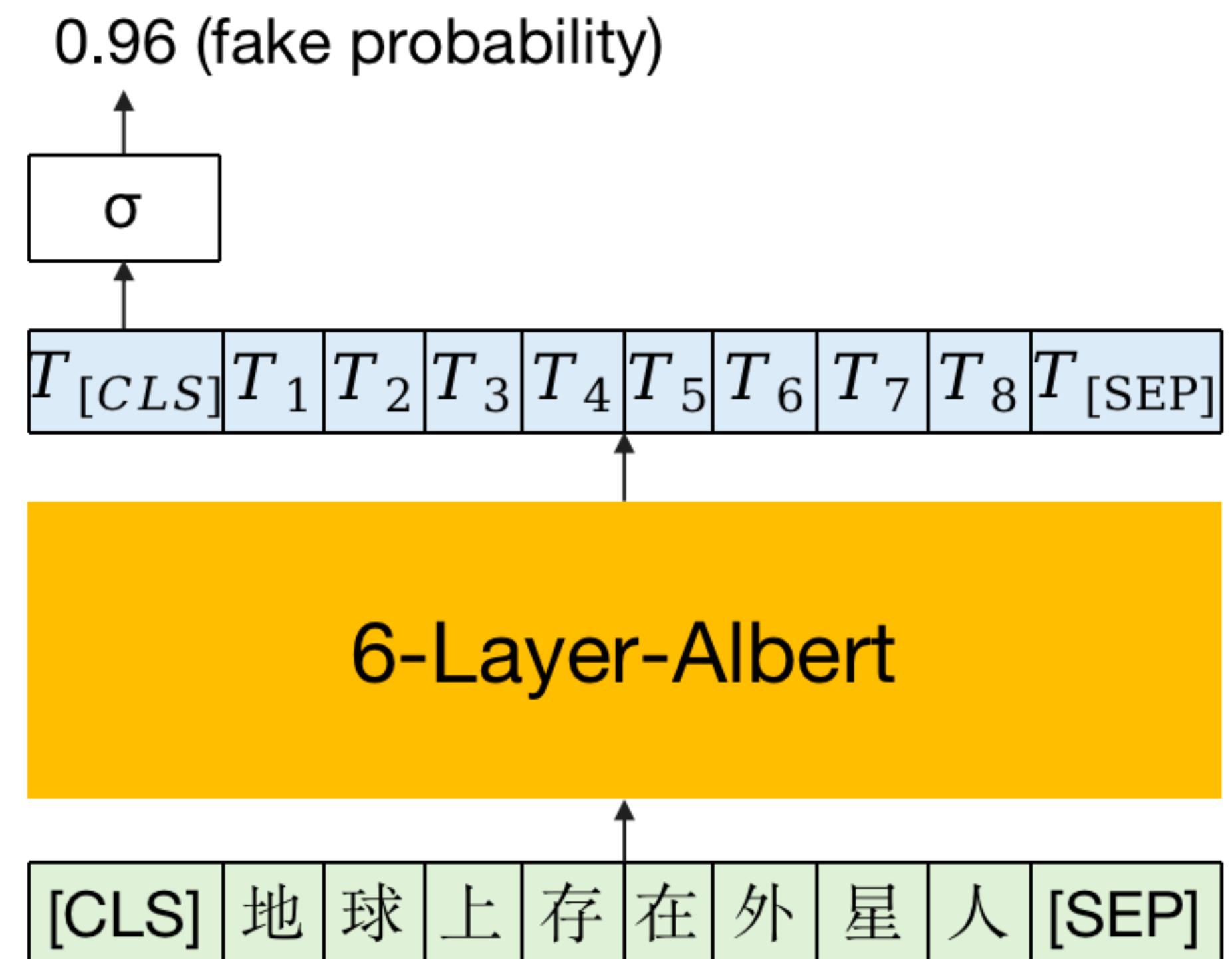
数据获取：从科学科技类历史打压队列中获取数据，**不需要额外进行人为标注**，打压的作为正例，全端推荐的作为负例

数据增强：正例重复采样，负例亚采样的方式，比例1:5左右

构造数据：按时间由远及近以10:1:1构造训练集、验证集和测试集

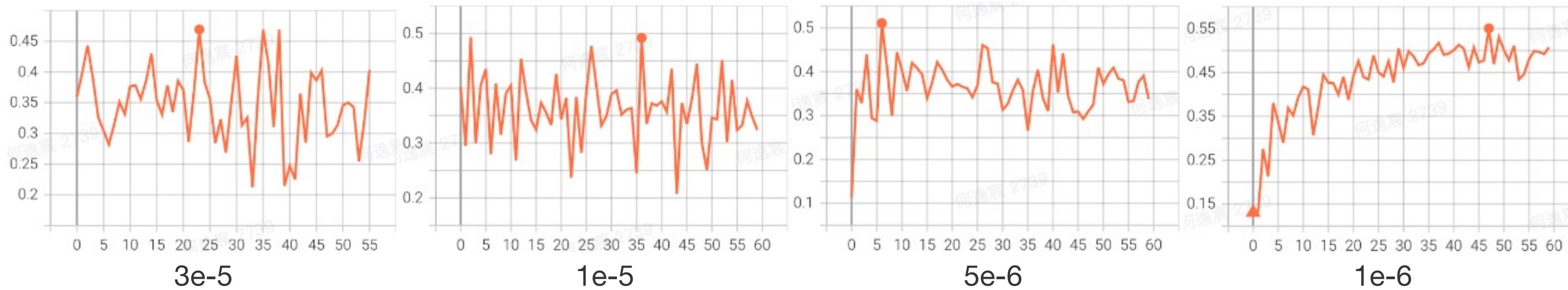
模型选取：如右图所示的6层Albert模型

人工检验：将模型命中的数据进行人工标注，验证模型的打压率



伪科学内容识别 — 技巧

1. 作者会根据被打压的情况改变写作风格，用时间越靠近的数据训练效果越好
2. 由于正负例不均衡，正例较少，采用正例双倍采样可以提升模型效果，**self-labeling**效果一般
3. 对于微头条和问答体裁，加入图文数据会显著提高模型效果，但会降低在图文测试集上的效果
4. 增加字体排版特征，取头尾各256词作为模型输入
5. 之前模型学习率设置不合理，通过实验发现 $1e-6$ 较合理



No	数据组成	测试集	P	R	F1
1	回答	回答	0.51	0.43	0.47
2	回答 + 正例double	回答	0.57	0.45	0.50
3	回答+图文	回答	0.70	0.53	0.60
4	图文	图文	0.66	0.77	0.71
5	微头条+图文	图文	0.63	0.56	0.59
6	图文	图文	0.60	0.87	0.71
7	图文+pseudo标签	图文	0.58	0.89	0.70
8	增加特殊字体token	图文	0.65	0.84	0.73
9	增加特殊字体token &首尾各256词	图文	0.69	0.83	0.76

伪科学内容识别 — 收益

1. 四体裁伪科学模型打压率由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，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. 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%
5. 协助处理伪科学发文账号220个，虚假内容录入谣言库共计170余条

测试集效果：

体裁	旧模型打压滤	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%



No.	作品	GID	展现	VV	点击率	消费时长(小时)	推荐/搜索状态	审核状态	发文方式
1	用中国知识改变人类对宇宙的认知 否定万有引力定律、否定太阳上是核反应挑战清华大学、北京大学、中科院大...	1720258116265998	2	0	0.00%	0	删除 原因：用户封禁	--	手工
2	用中国知识改变人类对宇宙的认知 否定万有引力定律、否定太阳上是核反应挑战清华大学、北京大学、中科院大...	1720184073265160	13	0	0.00%	0	删除 原因：用户封禁	--	手工
3	用中国知识改变人类对宇宙的认知 否定万有引力定律、否定太阳上是核反应挑战清华大学、北京大学、中科院大...	1720031582347399	0	0	--	0	删除 原因：用户封禁	--	手工
4	用中国知识改变人类对宇宙的认知 否定万有引力定律、否定太阳上是核反应挑战清华大学、北京大学、中科院大...	1719549616956429	95	0	0.00%	0	删除 原因：用户封禁	--	手工

01

虚假信息检测

02

伪科学内容识别

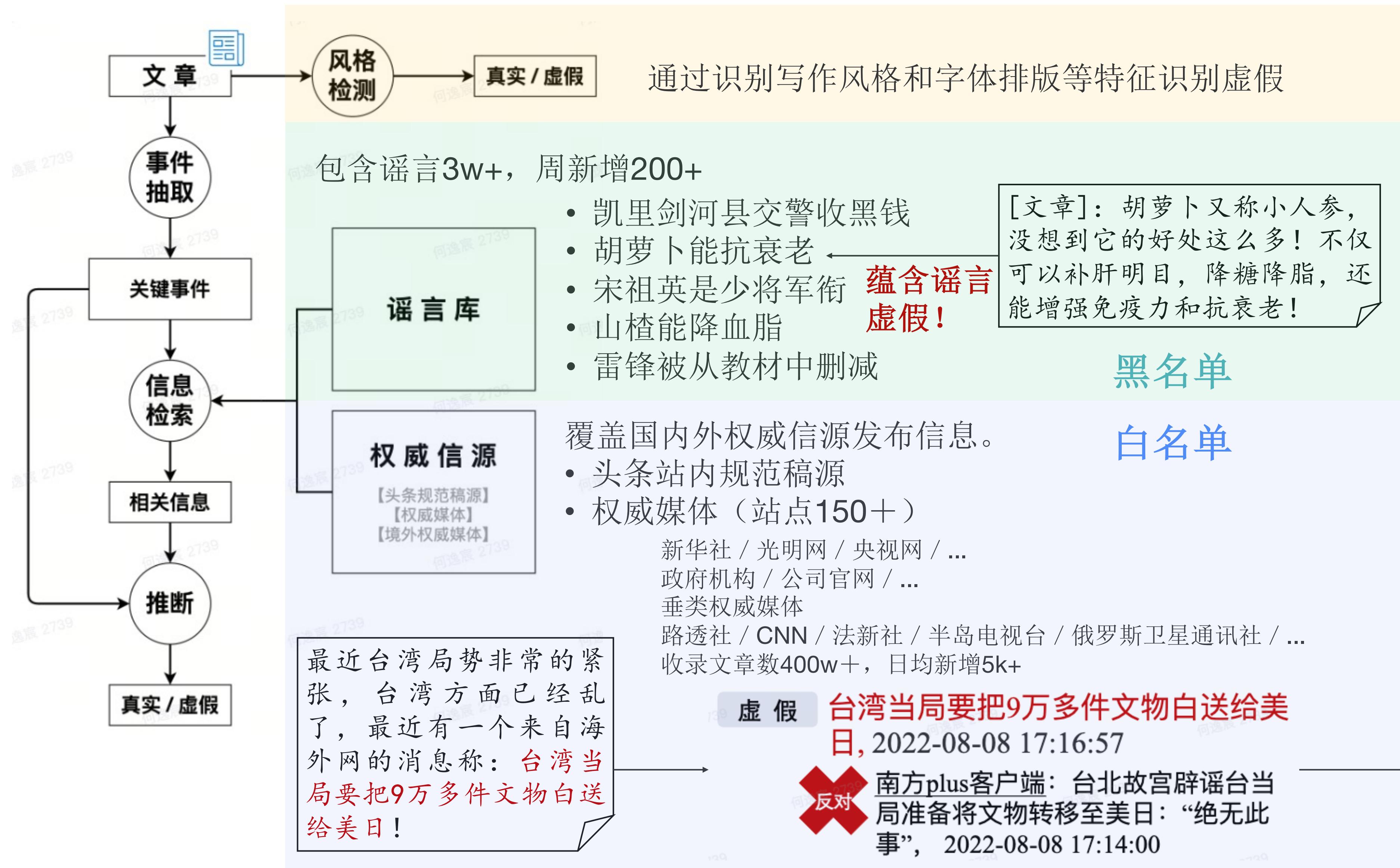
03

信源库系统搭建

04

通用审核模型构建

信源库系统搭建



本人从事的工作：

伪科学内容识别

信源库系统搭建

通用审核模型

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

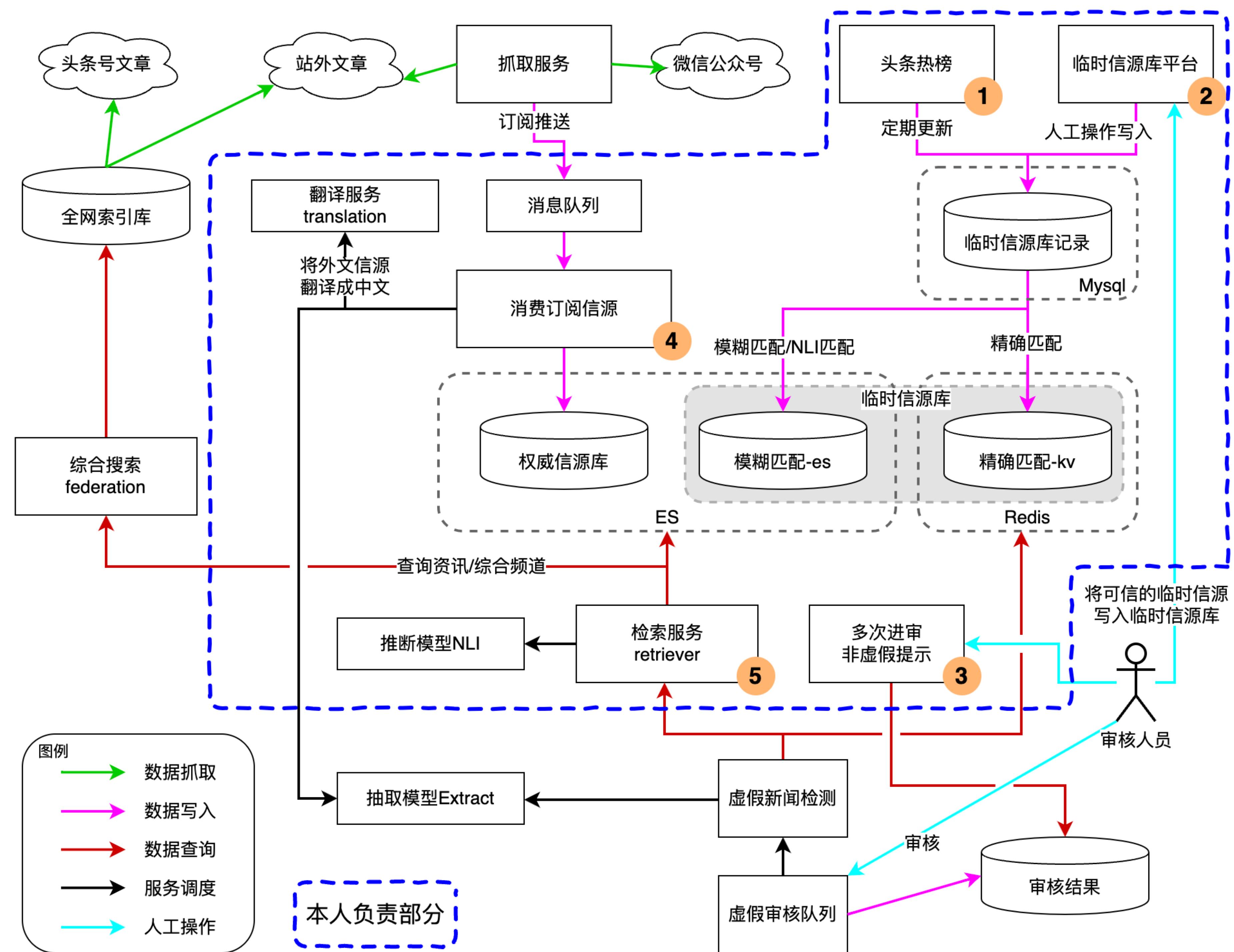
9万藏品将转移美日？台北故宫回应
鲁网 2022-08-08 16:08:30

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

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

#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

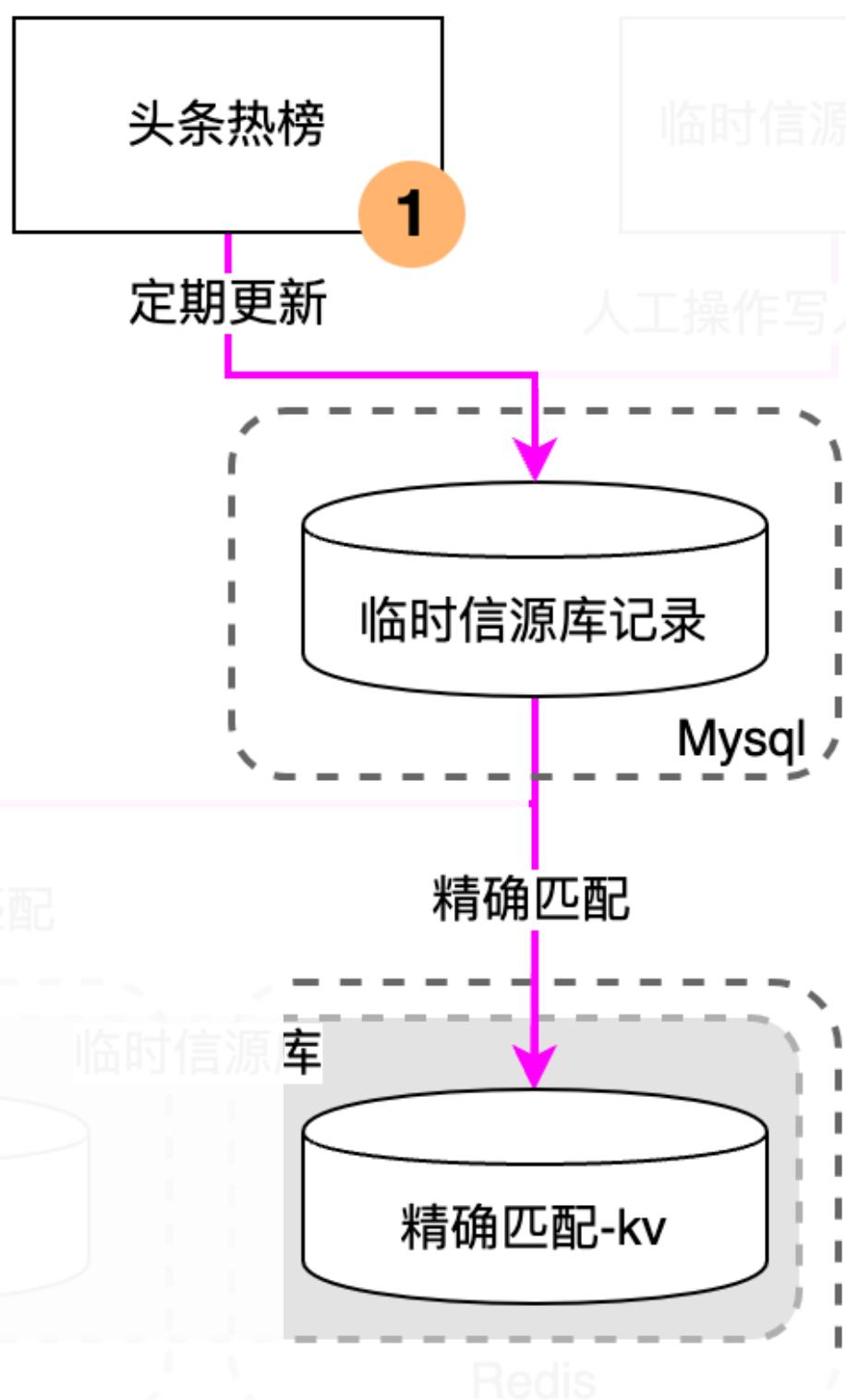
判断结果：无相关权威信源，疑似虚假

命中权威源 无

信息：

→ 服务调度
→ 人工操作

本人负责部分



p_date	进审量	命中热榜	热榜比例
2023-07-01	1,253	353	3.73%
2023-07-02	1,032	232	3.51%
2023-07-03	1,403	395	4.12%
2023-07-04	1,569	562	4.50%
2023-07-05	1,680	2,034	10.81%
2023-07-06	1,507	1,233	7.55%
2023-07-07	1,602	521	3.74%

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

通过临时信源库平台录入临时信源。

多次进审不打压的case视为候选临时信源，由人工判断决定是否将其加入临时信源库



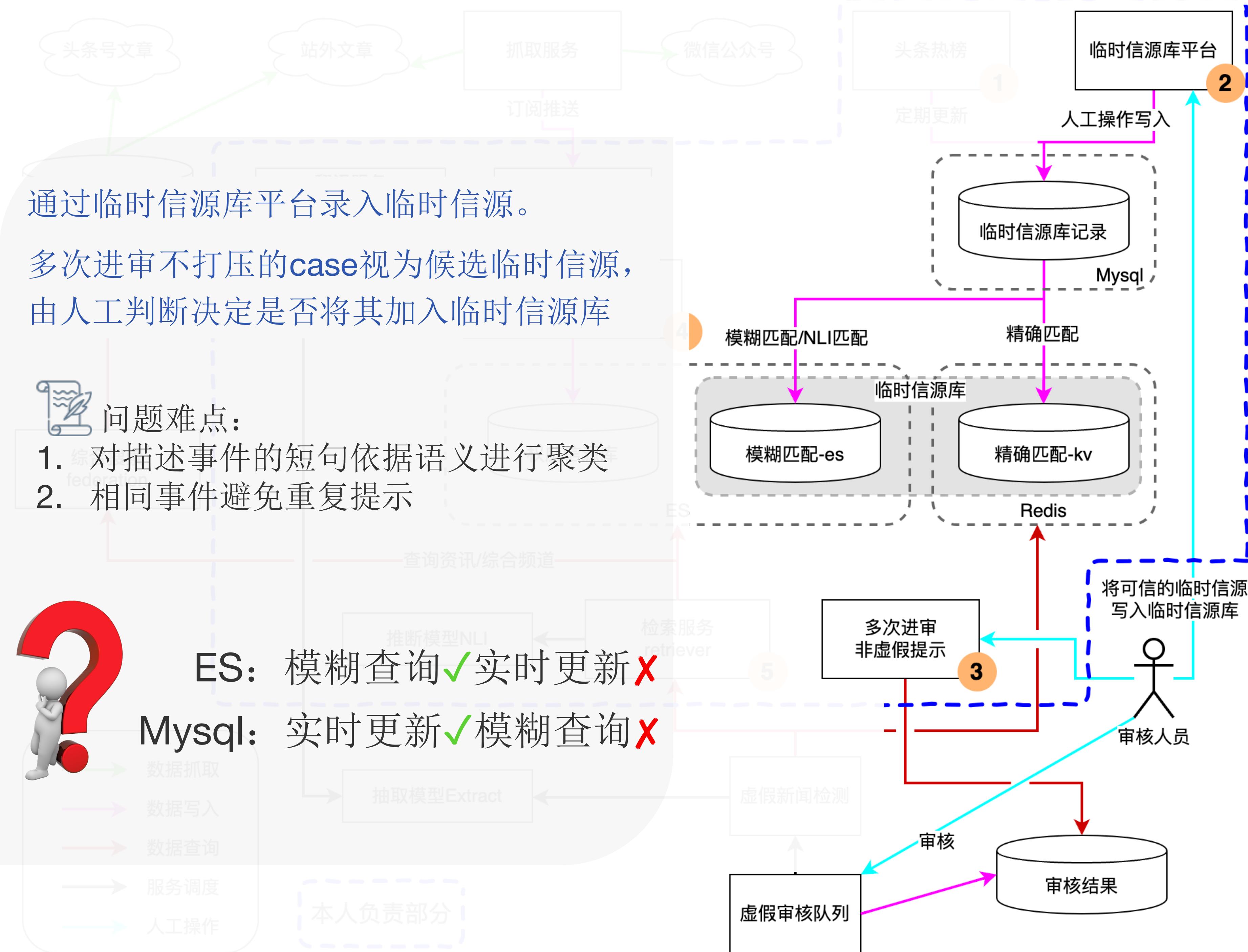
问题难点：

- 对描述事件的短句依据语义进行聚类
 - 相同事件避免重复提示

A 3D white humanoid figure with its hand to its chin, standing next to a large red question mark.

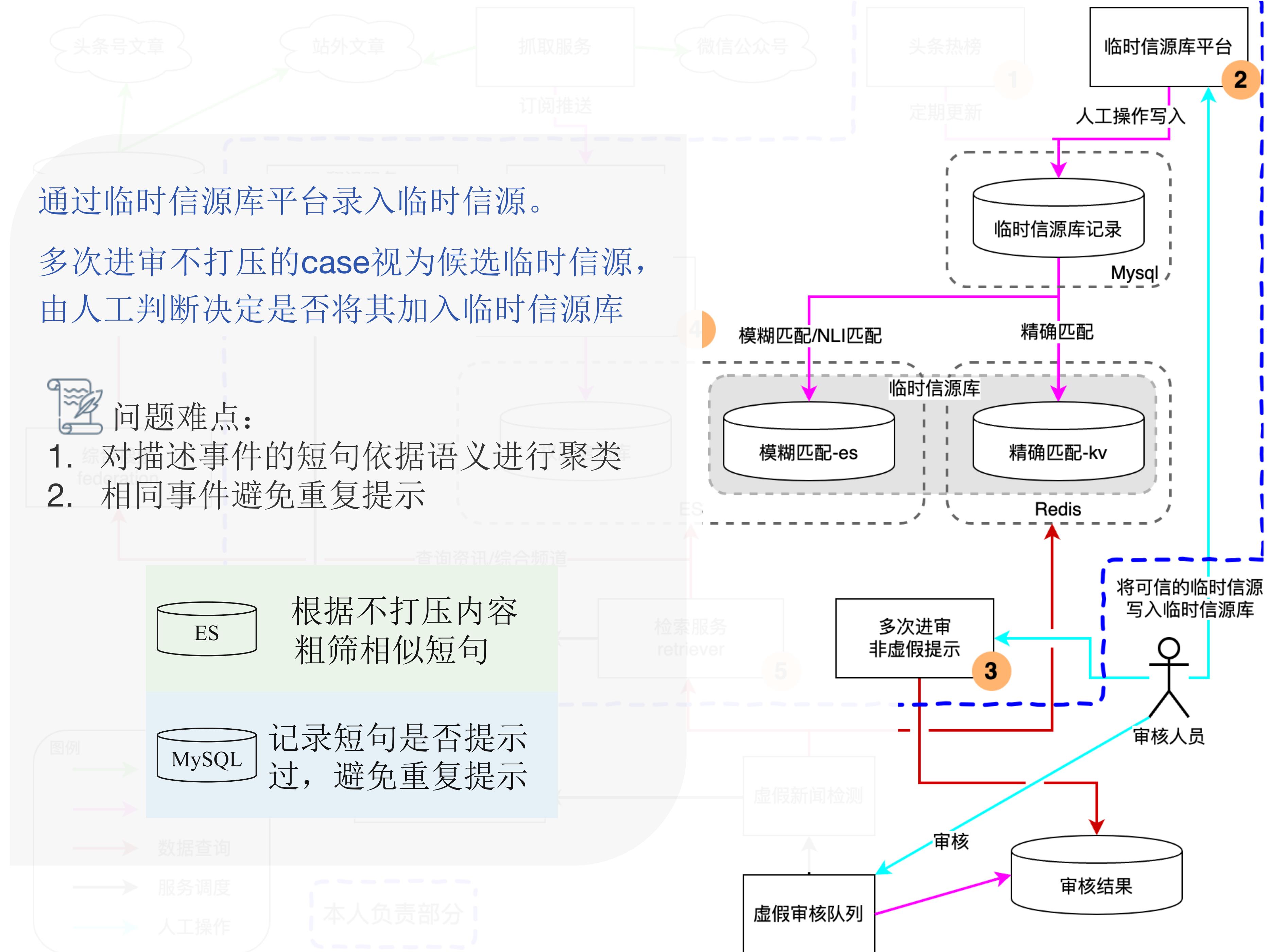
ES: 模糊查询 ✓ 实时更新 ✗

Mysql: 实时更新 ✓ 模糊查询 ✗



信源库系统搭建

- 热榜内容自动入库
- 临时信源库
- 多次进审非虚假提示
- 订阅信源消费
- 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

增

创建

* 事件内容

匹配方式 精确匹配 NLU匹配

创建人

过期时间 (默认一天有效)

创建 取消

虚假新闻临时信源提示群 16 | 2

临时信源

消息 Pin 群公告 云文档 +

虚假新闻bot 机器人 | 用于查询虚假新闻以及定位

打压数量: 1
美参议员: 如果中国封锁台湾, 美国就封锁中国 https://gip.bytedance.net/bluewhale/content/article_0
不打压数量: 4
美国南卡参议员林赛·格雷厄姆叫嚣: “美国需要中国明白, 如果他们封锁台湾, 美国就会封锁中国 https://gip.bytedance.net/bluewhale/content/article_gid=1762831653118988
美国南卡参议员林赛·格雷厄姆叫嚣: “美国需要中国明白, 如果他们封锁台湾, 美国就会封锁中国 https://gip.bytedance.net/bluewhale/content/article_gid=1762818800403468
美国南卡参议员林赛·格雷厄姆叫嚣: “美国需要中国明白, 如果他们封锁台湾, 美国就会封锁中国 https://gip.bytedance.net/bluewhale/content/article_gid=1762826507550735
美顶级智库: 中国大陆若武统台湾, 美国就封锁中国海空贸易通道 https://gip.bytedance.net/bluewhale/content/article_0

OK | 罗婧姝

查

输入关键词查找 刷新 请选择 ^ luojingshu 新建

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	NLU匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/7 00:00:00	编辑 下架
38479	美国政府周三表示, 他们将对尼古拉·尼古拉耶夫等委内瑞拉最高司法委员会的4名成员实施签证制裁, 理由是他们涉嫌“严重腐败”	NLU匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/6 09:59:59	编辑 下架
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	编辑 下架
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	NLU匹配	在线	2023/4/7 00:00:00	luojingshu	2023/4/5 14:06:23	编辑 下架
38310	西南大学一教授被曝潜规则女博士三年	NLU匹配	过期	2023/4/6 00:00:00	luojingshu	2023/4/5 13:53:03	编辑 下架

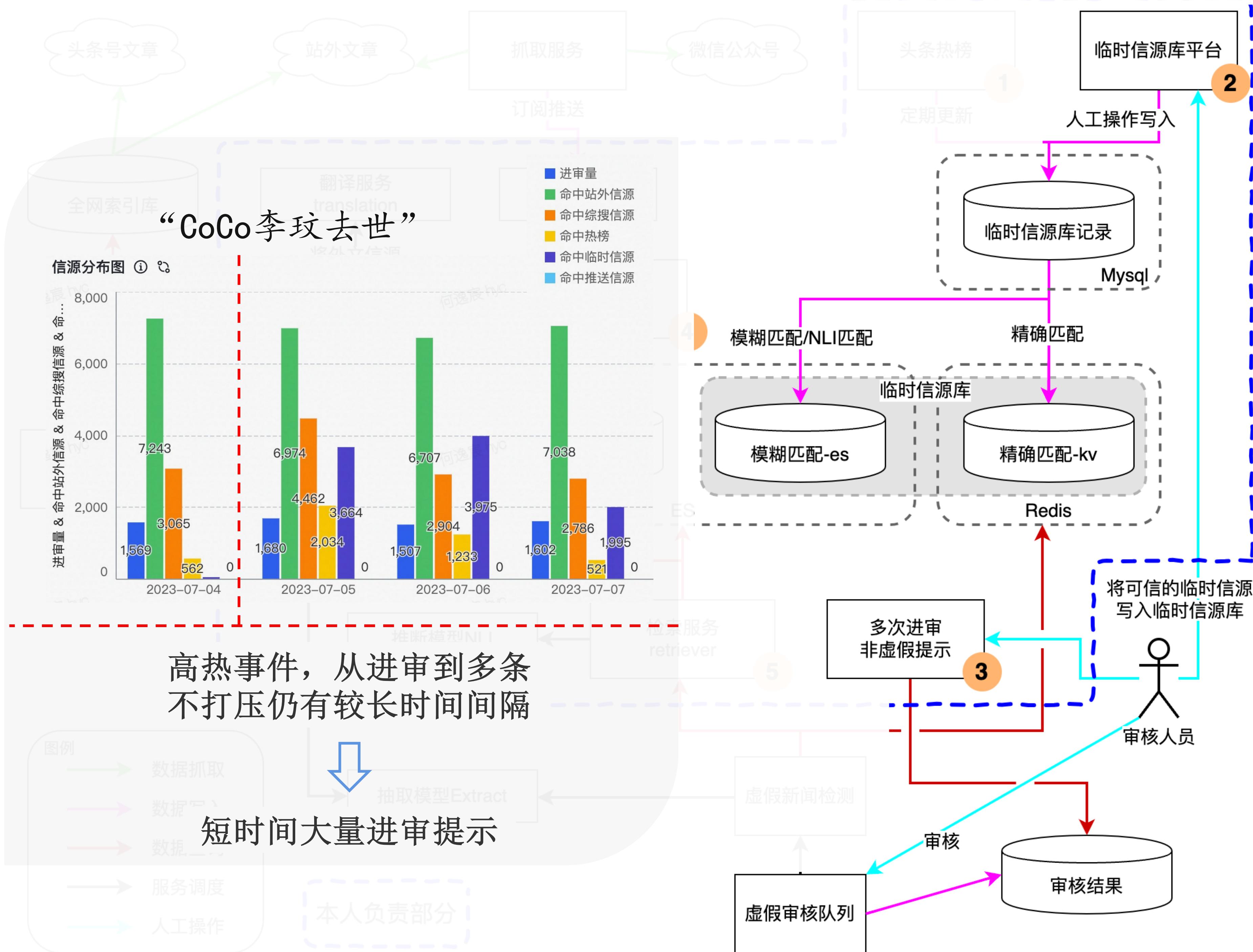
共 221 条 < 1 2 3 ... 23 >

改

删

信源库系统搭建

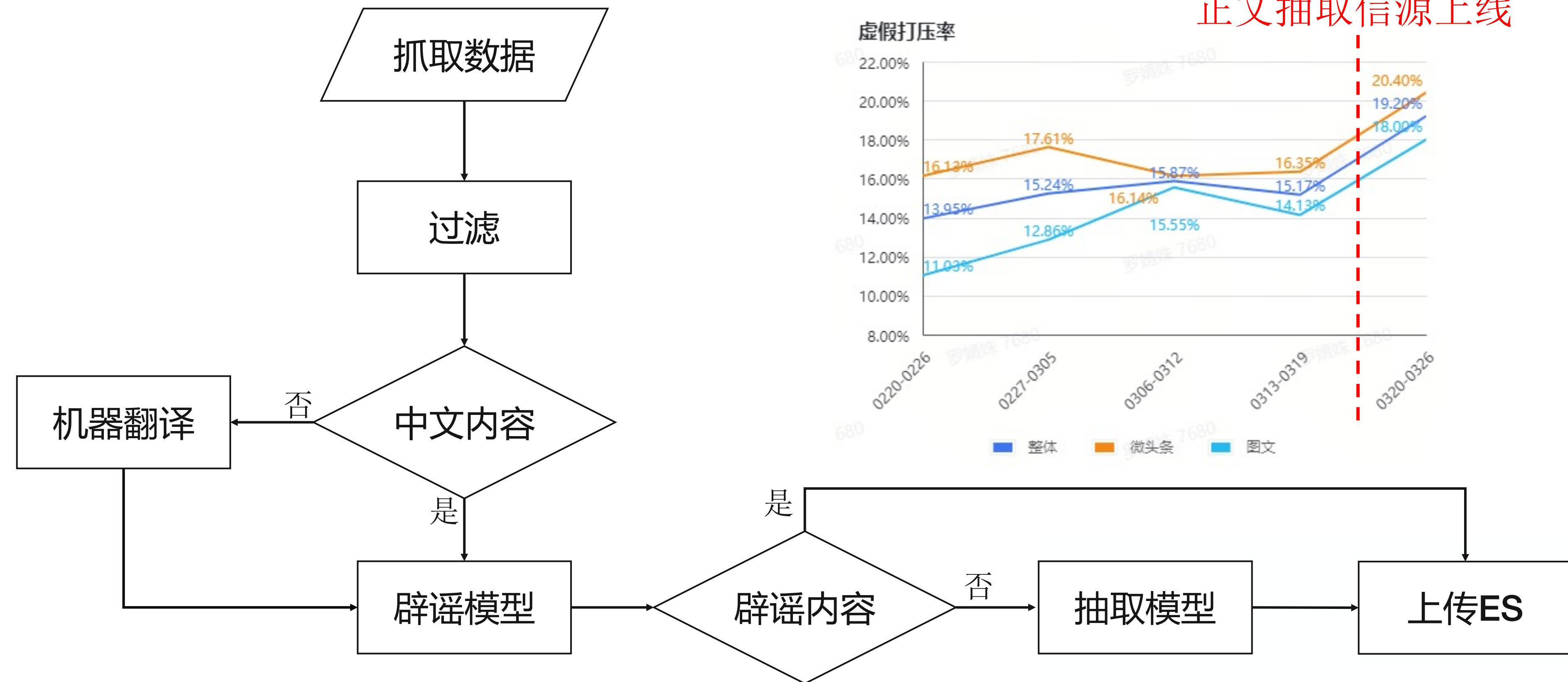
- 热榜内容自动入库
- 临时信源库
- 多次进审非虚假提示
- 订阅信源消费
- 检索服务



信源库系统搭建

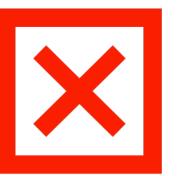
- 热榜内容自动入库
- 临时信源库
- 多次进审非虚假提示
- 订阅信源消费
- 检索服务

台当局要把文物送给美日，台北故宫博物院辟谣，绝无此事



消费舆情中台抓取并写入kafka队列的信源，
将其处理成特带数据结构并写入权威信源库。

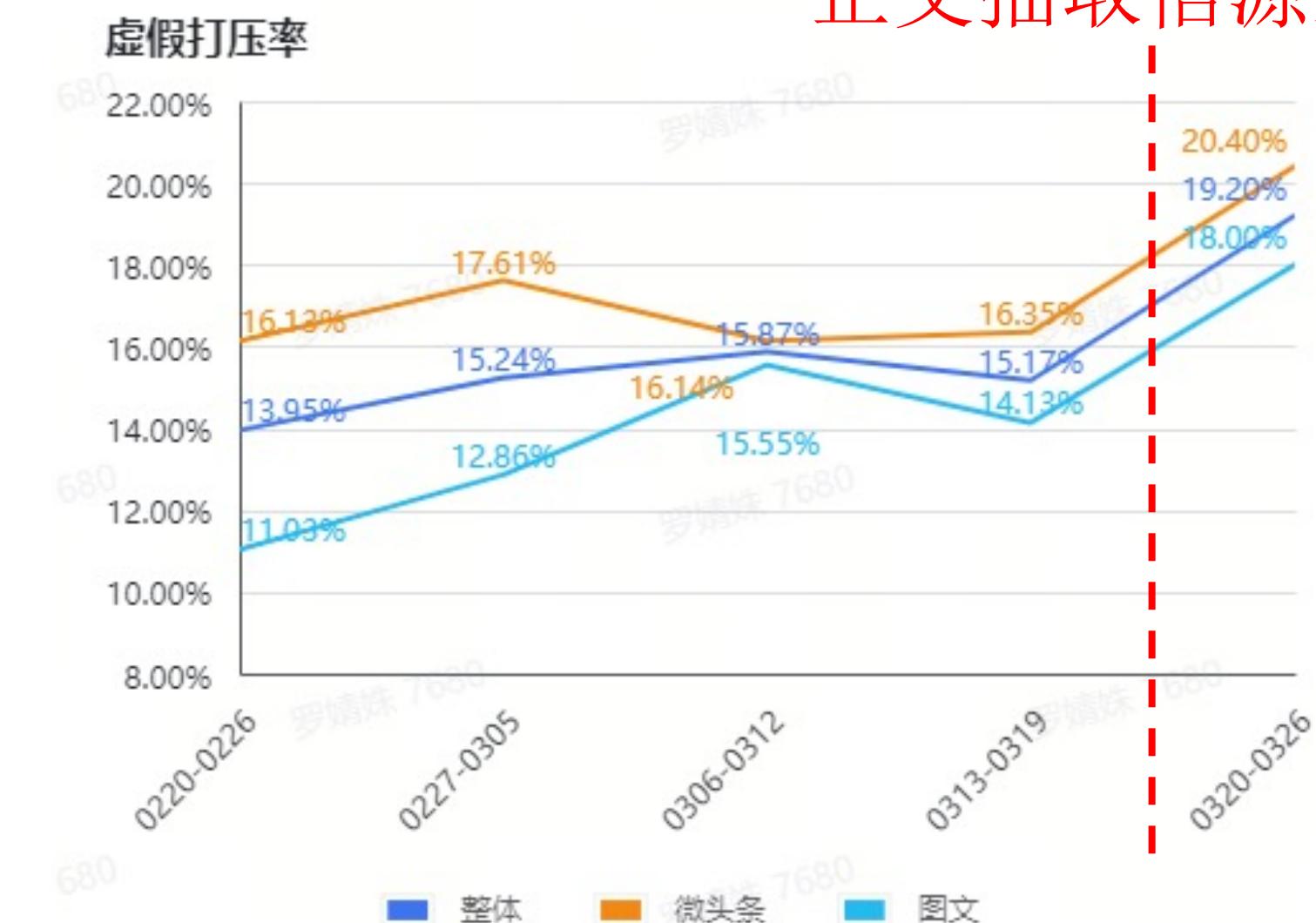
消费延迟
只有标题作为信源



Golang重构
抽取正文、删除辟谣



正文抽取信源上线

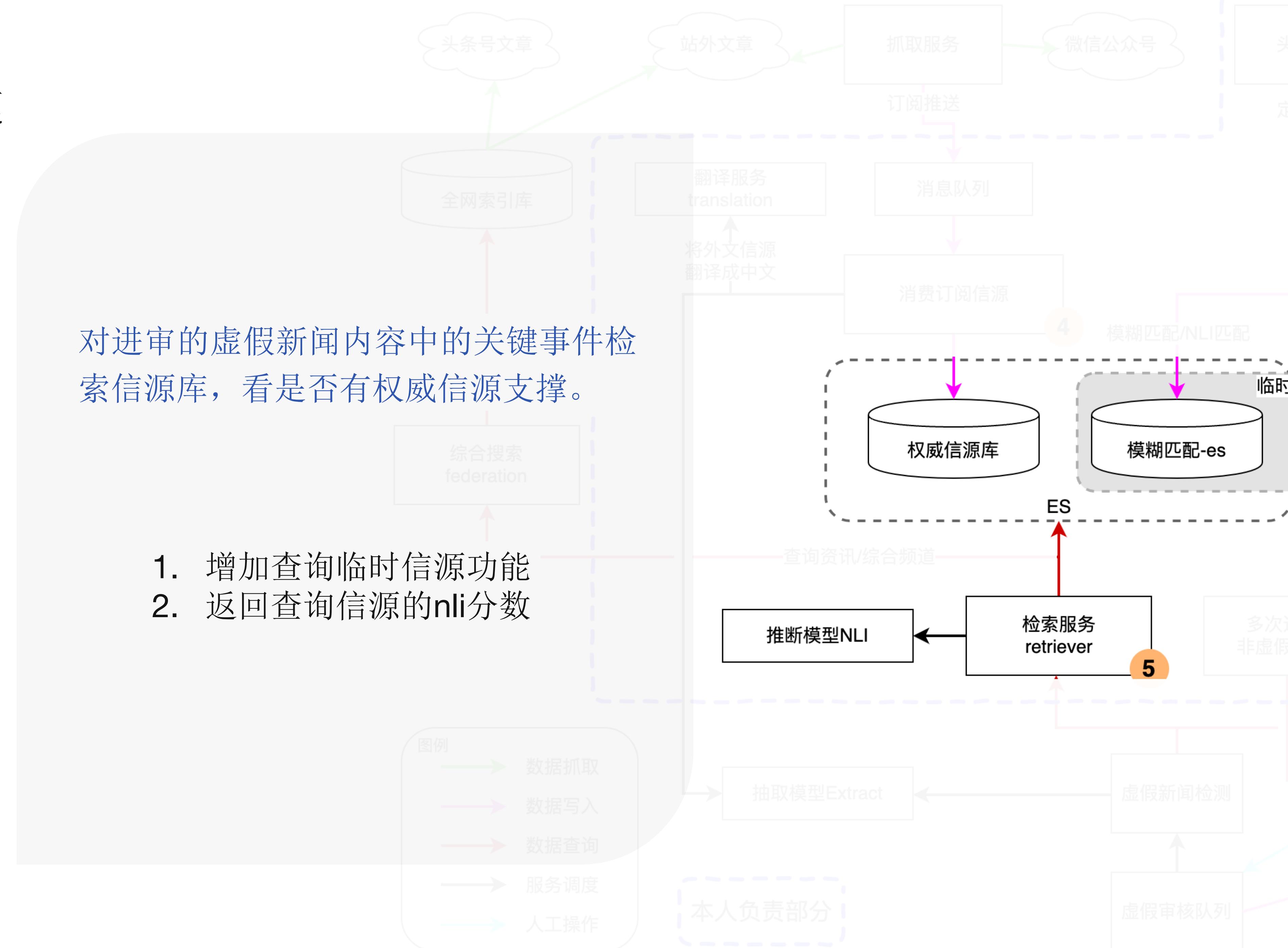


信源库系统搭建

- 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

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。

1. 增加查询临时信源功能
2. 返回查询信源的nli分数



01

虚假信息检测

02

伪科学内容识别

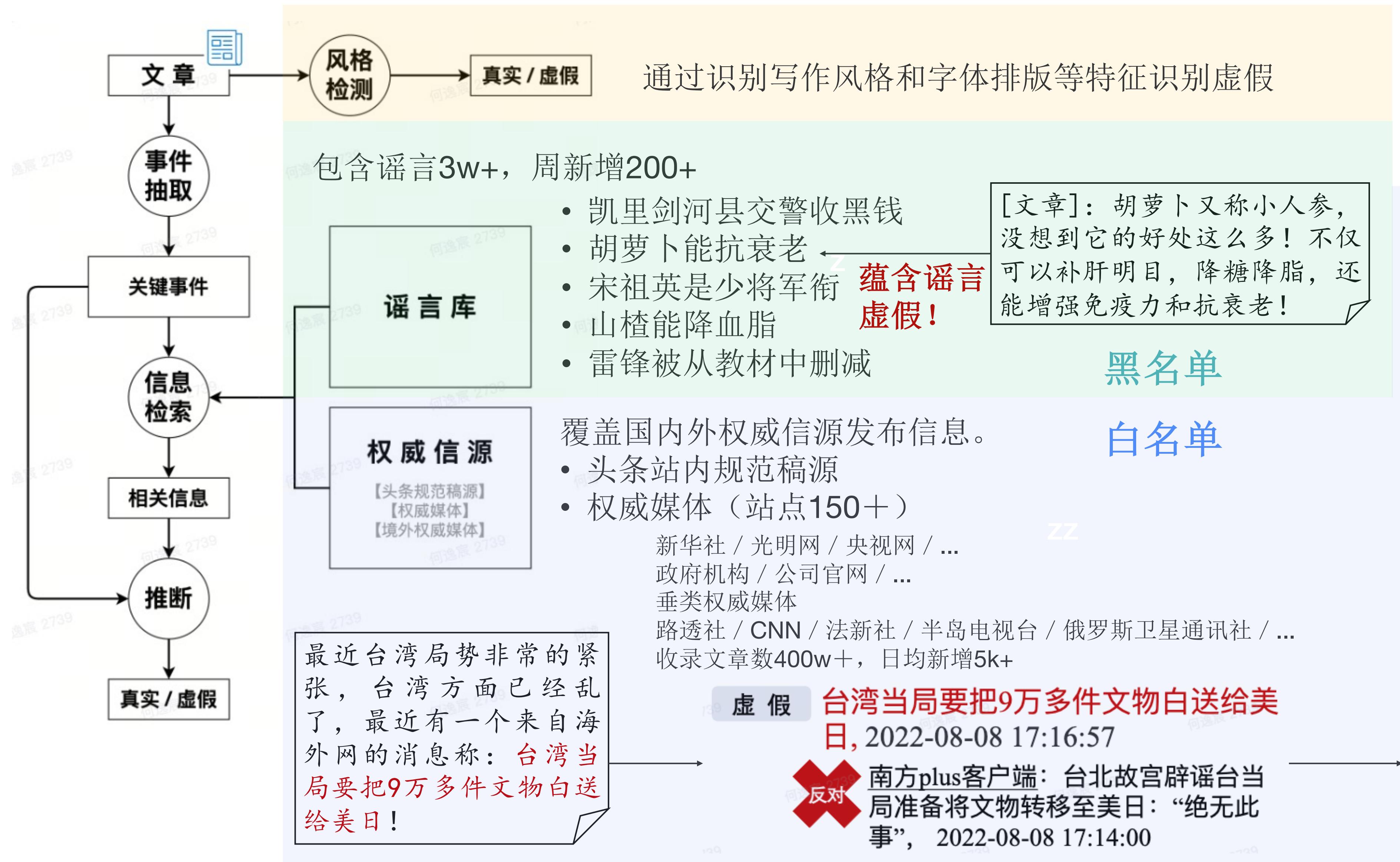
03

信源库系统搭建

04

通用审核模型构建

通用审核模型



本人从事的工作:

伪科学内容识别

信源库系统搭建

通用审核模型

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

通用审核模型 — 业务场景大模型训练流程

特点和优势

1. 总结和推理能力 ⇒ 总结文章内容并和谣言点、审核政策等短句进行匹配
2. 学习能力 ⇒ 学到训练集中的审核尺度和规则

使用方法

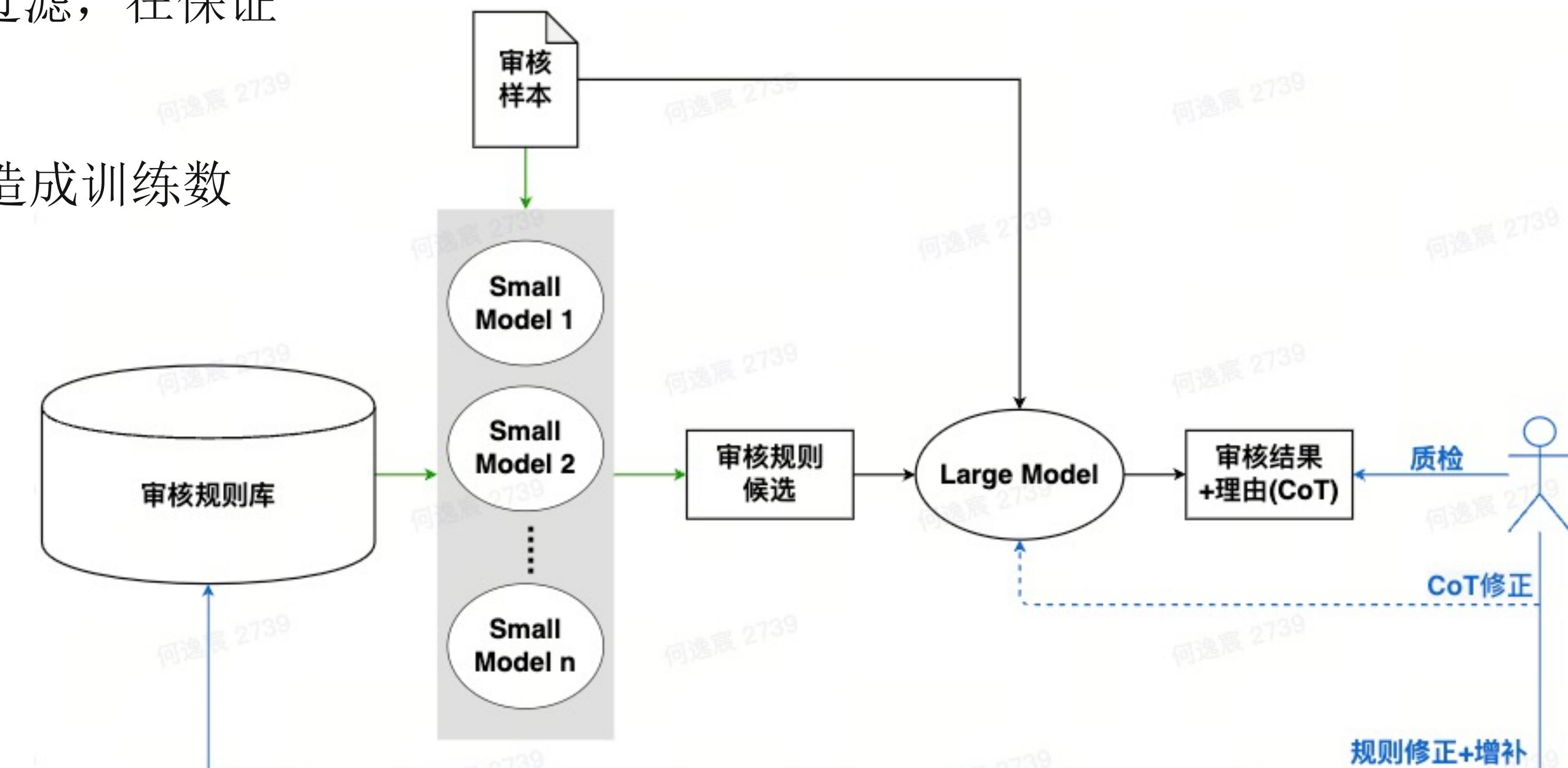
1. 在小模型召回的基础上，增加大模型作为过滤，在保证尽量不漏召的情况下，提高模型的准确率。
2. 在日常的迭代中，将每周误提示的case构造成训练数据，并定期更新模型。

目标

1. 提高准确率和召回率
2. 替代人工审核

模型输入

- [文章]: 一段待审核的文本
[句子]: 谣言点、审核政策等
[规则]: 额外的补充规则



通用审核模型 — 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: 支持中文输入的增强版LLaMA，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进行*continue pretrain*。使用40B的平行语料和中文单语语料进行了预训练，扩展词表，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，词表大小为10万，使用250B tokens进行训练。

vicuna: 通过使用从ShareGPT.com获取的约7万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、13B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，推理速度较快。

LLaMA 2、SEED等

通用审核模型 — 数据集

场景	数量	prompt
知识库	4k	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”。
审核政策	6k	在审核政策识别场景下。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。
审核政策nli	264k	针对审核政策识别任务，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。
新闻虚假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。

通用审核模型 — 训练数据构造

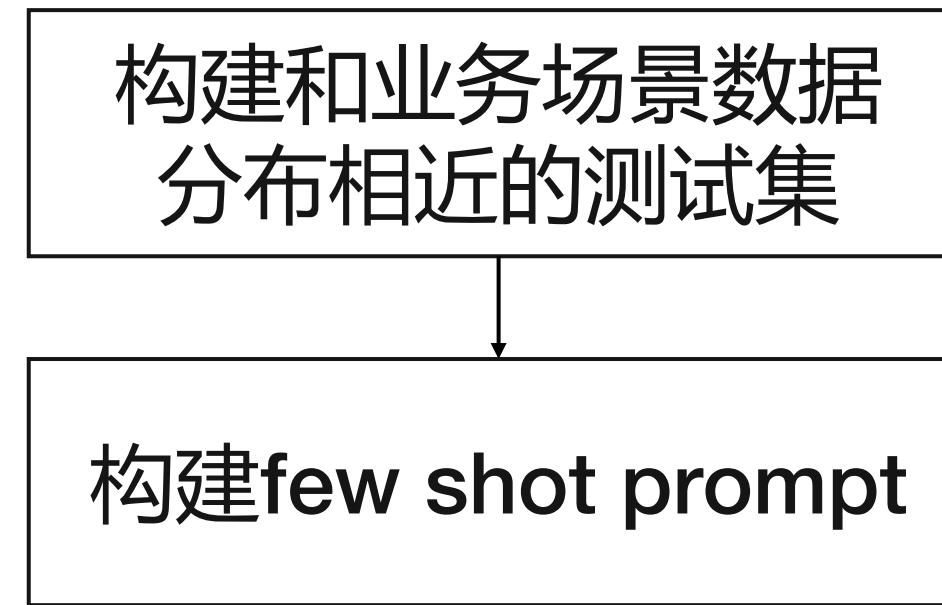
构建和业务场景数据
分布相近的测试集

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。

数据分布：

- 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)

通用审核模型 — 训练数据构造



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。

数据分布：

- 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

通用审核模型 — 训练数据构造

few shot prompt构造

针对易错负例构造 主体、时态和矛盾等

请判断[文章]是否蕴含[句子]，请回答是或否，并说明理由

[文章]事业编制改革，大批事业编制要被取消，医生与老师首当其冲！事业单位迎来动态普调，有42种事业编制被取消，改为“聘任制”，这些事业单位的工作人员将与单位签订劳动合同，合同到期之后，根据双方意愿选择是否继续续签。

[句子]国家全面取消事业编制。

否。[文章]表述内容为“部分岗位取消事业编制”，而[句子]表述为国家层面事业编制全部被取消，该说法过于绝对，和[文章]表述**主体范围**不同。

[文章]政协委员建议把改“六三三”学制为“五二三”，并将义务教育由九年改为十年。相应地，可取消中考普职分流，多数初中生毕业后直接进入普通高中，形成小学、初中、普通高中贯通办学体制。

[句子]普职分流取消。

否。[文章]表述内容为建议取消普职分流，该情况属实，而[句子]表述为完成时，已经取消普职分流，[句子]和[文章]表述**时态**不同。

[文章]丁女士点点头，有些犹豫的问道：“听说孕吐和宝宝的智商有关，孕吐的越厉害宝宝的智商越高，这是真的吗？”丁女士的丈夫还接道听说不孕吐生出来的宝宝智商不咋地。我笑了笑点点头说道，“确实听到过这样的说法，但是这没有一点科学依据，其实孕吐的反应的轻重不能决定宝宝的智商高低。即使有的孕妇在孕期吐得很严重，生出的宝宝智商很高，那是一种巧合。”

[句子]孕吐越严重孩子越聪明

否。[文章]表述孕吐越严重孩子越聪明没有一点科学依据，是对[句子]表述的**辟谣**。

通用审核模型 — 训练数据构造

few shot prompt构造

使用 COT: 总结 + 判断 能提高模型效果

p/r/f: 0.31 0.91 0.46

p/r/f: 0.37 0.84 0.51

[文章]事业编制改革，大批事业编制要被取消，医生与老师首当其冲！\n事业单位迎来动态普调，有42种事业编制被取消，改为“聘任制”，这些事业单位的工作人员将与单位签订劳动合同，合同到期之后，根据双方意愿选择是否继续续签。

[句子]国家全面取消事业编制。

1. [文章]提到有42种事业编制被取消。2. [句子]提到国家将全面取消事业编制。3. 42种事业编制被取消不能说明全面取消事业编制。4. [文章]不能推出[句子]，否。

[文章]政协委员建议把改“六三三”学制为“五二三”，并将义务教育由九年改为十年。相应地，可取消中考普职分流，多数初中生毕业后直接进入普通高中，形成小学、初中、普通高中贯通办学体制。

[句子]普职分流取消。

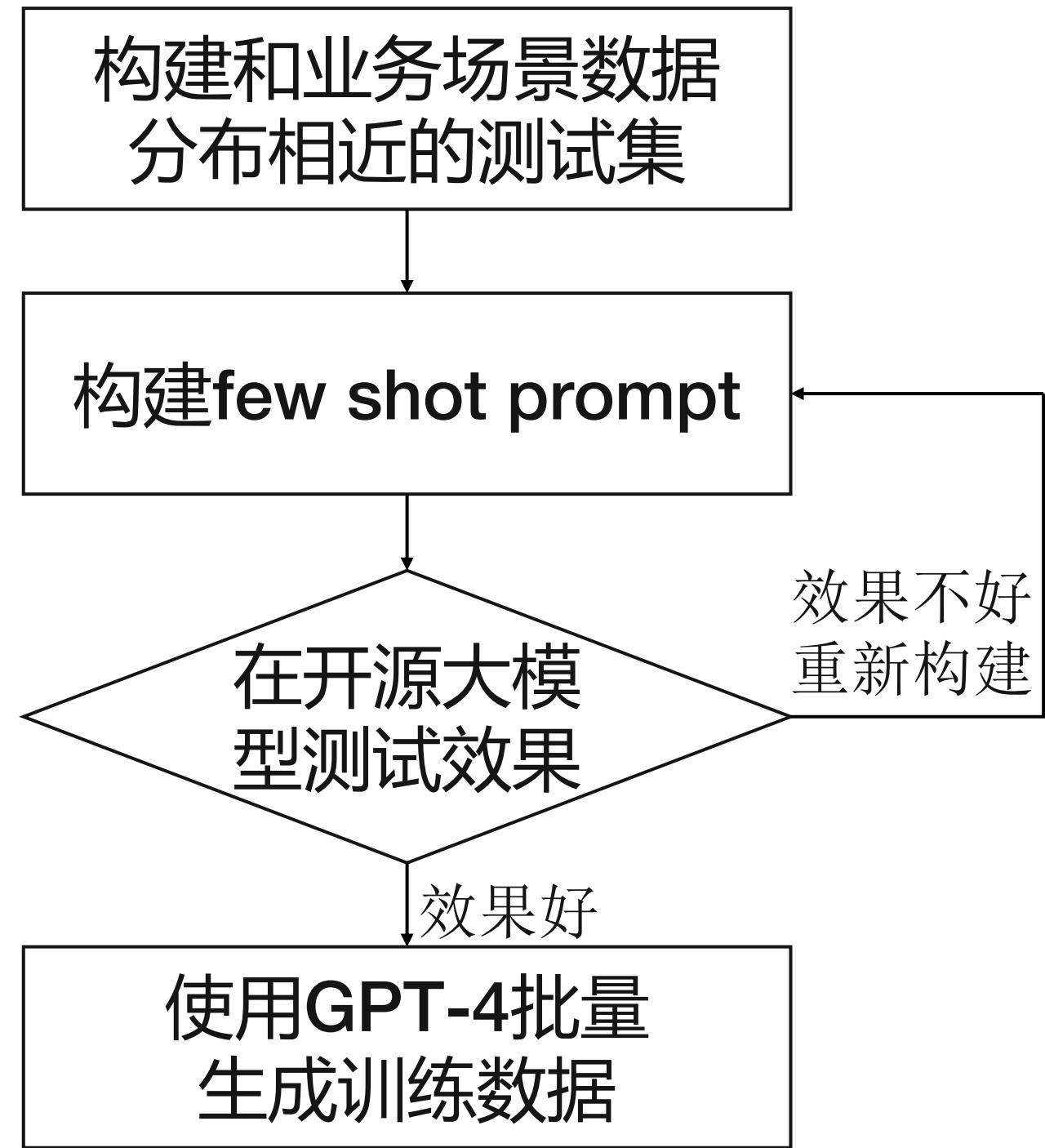
1. [文章]提到政协委员建议取消中考普职分流。2. [句子]提到普职分流取消。3. 政协委员建议取消普职分流是政协委员本人的观点，是假设和虚构的，而普职分流取消是事实，观点并不能推出事实。4. [文章]不能推出[句子]，否。

[文章]丁女士点点头，有些犹豫的问道：“听说孕吐和宝宝的智商有关，孕吐的越厉害宝宝的智商越高，这是真的吗？”丁女士的丈夫还接道听说不孕吐生出来的宝宝智商不咋地。我笑了笑点点头说道，“确实听到过这样的说法，但是这没有一点科学依据，其实孕吐的反应的轻重不能决定宝宝的智商高低。即使有的孕妇在孕期吐得很严重，生出的宝宝智商很高，那是一种巧合。”

[句子]孕吐越严重孩子越聪明

1. [文章]提到孕吐越严重孩子越聪明没有一点科学依据，说明[文章]反对孕吐越严重孩子越聪明。2. [句子]提到孕吐越严重孩子越聪明，说明句子支持孕吐越严重孩子越聪明。3. 所以[文章]和[句子]呈相反意见，是对[句子]的辟谣。4. [文章]不能推出[句子]，否。

通用审核模型 — 训练数据构造



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。

数据分布：

- 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

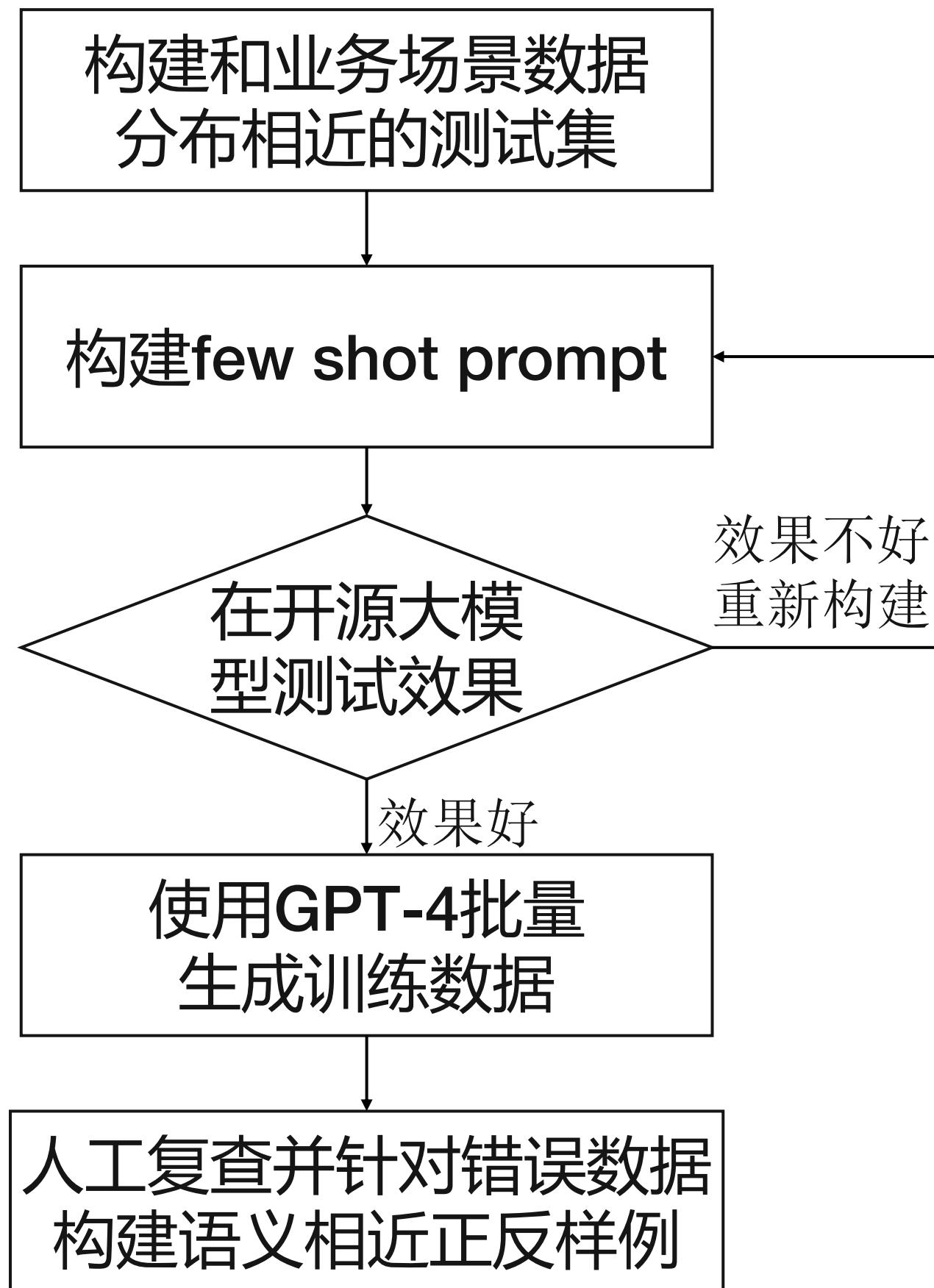
判断依据：

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。

通用审核模型 — 训练数据构造

帮助模型更好地理解业务判断尺度

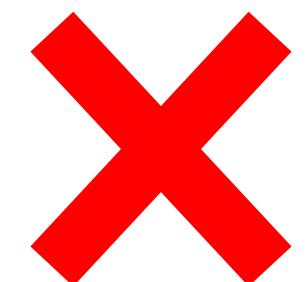
提升解决疑难问题的能力



【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。

通用审核模型 — 实验结果

训练数据和模型参数量的影响

No	model	precision	recall	F1 score
1	GPT4 5 prompt	0.70	0.90	0.788
2	llama30b 170 prompt	0.67	0.63	0.649
3	llama13b 1k prompt	0.68	0.68	0.676
4	llama13b 2k prompt	0.63	0.87	0.732
5	llama30b 2k prompt	0.66	0.90	0.763
6	llama30b 4k prompt	0.70	0.84	0.766
7	llama65b 4k prompt	0.69	0.89	0.777
8	llama-plus13b 0505 4k prompt	0.73	0.80	0.760
9	gpt2lab 13b 4k prompt	0.72	0.79	0.756
10	gpt2lab 13b 2k high quality prompt	0.71	0.87	0.782

模型效果的提升
得益于
1. 训练数据量增加
2. 模型参数量增加
3. 提高训练数据的质量

通用审核模型 — 实验结果

语言、数据构成的影响

No	model	precision	recall	F1 score	备注
1	online albert (中)	0.56	0.96	0.706	
2	GPT4 few shot (中)	0.70	0.90	0.788	
3	llama13b (英)	0.70	0.77	0.733	谣言库训练数据 翻译成英文
4	llama30b (英)	0.72	0.82	0.769	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。
5	llama65b (英)	0.72	0.84	0.777	
6	llama-plus 13b (中)	0.73	0.80	0.760	中文预训练模型
7	gpt2lab 13b (中)	0.72	0.79	0.756	谣言库训练数据
8	llama-plus 13b (中)	0.72	0.84	0.777	谣言库+审核政策+写作数据 训练数据
9	gpt2lab 13b (中)	0.72	0.83	0.772	将谣言库和审核政策训练数据的prompt进行多样化改写，提升模型对prompt的理解能力
10	llama-plus 65b (中)	0.77	0.87	0.816	

通用审核模型 — 实验结果

理由和结论的位置

1. 在训练中使用COT能提高最终的效果
2. 测试中是否使用COT效果变化不明显
3. 使用先结论的prompt可以在实际应用中提高模型推理速度

是。理由：1. [文章]提到半夜3-5点是肺经当令的时候，肺气血最为旺盛，可以濡养我们的形体官窍。2. [句子]提到x点到x点是身体造血的时候。3. 肺气血旺盛可以濡养形体官窍可以理解为身体造血的时候。

是。理由：1. [文章]提到如果把英语降为副科，比如分值降到60分。2. [句子]提到英语降为副科。3. 把英语降为副科是[文章]中的一种假设。

★ 线上模型提问方式

请判断[文章]是否能推出[句子]，先解释原因，然后回答“是”或“否”。**先理由后结论**

请判断[文章]是否能推出[句子]，给出是或否的判断，并说明理由。**先结论后理由**

请判断[文章]是否能推出[句子]，回答是或否。**只给结论**

请判断这篇[文章]是否可以从语义上推出[句子]，如果可以，请解释原因并回答“是”；如果不行，请解释原因并回答“否”。

[文章]可以推出[句子]吗？请说明原因，并给出是或否的结论。

分析这篇[文章]能推导出[句子]所表达的意思吗？请说明理由，并在最后给出是或否的结论。

[文章]能推出[句子]吗？请说明理由，如果可以，请回答是；如果不行，请回答否。

请判断这篇[文章]是否能够推导出这个[句子]，如果可以，请解释原因并回答“是”；如果不行，请解释原因并回答“否”。

你能否从这篇[文章]中推出这个[句子]？请先解释原因，然后回答“是”或“否”。

请您分析一下这篇[文章]可否从语义上推出[句子]的意思，如果可以请回答是，如果不行请回答否；并给出判断的理由。

请问，[文章]可以推出[句子]吗？，给出是或否的结论并说明理由。

麻烦判断一下，这篇[文章]能推导出[句子]所表达的意思吗？给出是或否的结论并说明理由。

[文章]能推出[句子]的含义吗？如果可以，请回答是；如果不行，请回答否。并说明理由。

根据这篇[文章]，能否推出这个[句子]的意思？请给出判断并说明理由。

这篇[文章]是否足以从语义上推出这个[句子]的含义？请回答是或否，并解释理由。

通用审核模型 — 实验结果

65b模型的外部知识



[文章]在今年的两会期间，他再次提交了这一提案，但就像过去三年一样，这一提案今年被否决了。我真的不明白为什么全民免费医疗这么难。可以合理地说，我国的经济规模也很大，仅次于美国，在世界上排名第二。其他几个**金砖国家**，经济规模远不如我们，但他们实现了全民免费医疗，所以这么大的经济规模应该更容易实现！但是全民免费医疗直到现在才是空话。看病难看病贵已经是心照不宣的事了。每个人都应该去过医院。不用说，只有注册费是十几块钱开始的。这个价格只能挂在普通医生的号码上。专家号比较贵，往往几百。更可恨的是，有些专家还是被黄牛炒了，一个专家号可以炒到几万。

[句子]印度实施免费医疗

13b

1. [文章]提到和其他几个金砖国家实现了全民免费医疗，但没有具体提到印度。
2. [句子]提到印度实施免费医疗。
3. 由于[文章]没有提到印度实施免费医疗这一事实。因此，[文章]不能推出[句子]，否。

65b

1. [文章]提到和其他几个金砖国家一样，经济规模远不如我们，但他们实现了全民免费医疗。
2. [句子]提到印度实施免费医疗。
3. **金砖国家包括印度**。因此，[文章]可以推出[句子]，是。

通用审核模型 — 实验结果

65b模型的中英对照能力



[文章]缺乏灵活性：他们难以适应变化，并且拒绝调整业务战略。领导能力差：他们很难有效地激励和领导团队。同时，ChatGPT还列出了它眼中全世界最厉害的创业者。杰夫·贝佐斯（JeffBezos）：亚马逊的创始人，全球最大的在线零售商。埃隆·马斯克（ElonMusk）：SpaceX、Tesla、Neuralink和TheBoringCompany等公司的创始人。马克·扎克伯格（MarkElliotZuckerberg）：全球最大的社交网络平台Facebook的联合创始人。马云：阿里巴巴集团创始人，全球最大的电子商务公司之一。史蒂夫·乔布斯(SteveJobs)：苹果公司的联合创始人，

[句子]特斯拉由埃隆·马斯克创建

13b

1. [文章]提到埃隆·马斯克（ElonMusk）是SpaceX、Tesla、Neuralink和TheBoringCompany等公司的创始人。
2. [句子]提到特斯拉由埃隆·马斯克创建。
3. [文章]中并未提到特斯拉是埃隆·马斯克创建的。因此，[文章]不能推出[句子]，否。

65b

1. [文章]提到埃隆·马斯克是SpaceX、Tesla、Neuralink和TheBoringCompany等公司的创始人。
2. [句子]提到特斯拉由埃隆·马斯克创建。
3. Tesla即特斯拉，埃隆·马斯克是Tesla的创始人。因此，[文章]可以推出[句子]，是。

通用审核模型 — 收益和改进

收益

	召回量	准确量	准确率	覆盖率
nli	100	59	59%	100%
nli+llm	66	56	85%	95%

- 原提问方式不利于模型输出与训练集对齐
- 不同谣言点的打压尺度无统一标准



- 采用“总结-分析-结论”三段式提问，帮助模型输出对齐
- 补充规则字段，辅助模型进行判断

根据给定[文章]、[句子]和一些补充[规则]回答下列[问题]。

[文章]：于人体，可以杀灭水蛭，蚯蚓，蛔虫……苦棟子的出油率很高，因此也用来榨油，制作肥皂，油漆等。冬季寒凉，皮肤易生冻疮，因为油脂丰富，老一辈们也会去采摘一些苦棟子，去核留出皮肉，把它做成油膏，在冻疮上面一抹，涂抹几次冻疮即可痊愈。苦棟子有清热燥湿，行气止痛的功效，煮水可用于胃痛、肝气不舒所致的胸胁疼痛。由于其煎液还有较强的杀菌作用，因此对于皮肤病、头癬也有一定的抗菌止痒的效果。苦棟树还有一个用处，那就是制作装饰品，将其串成手串，拿在手里盘玩，极具特色，曾经也是很多文玩朋友的心爱之物。“棟花香，棟果苦”。苦棟树的花，叶，果，皮，根皆可入药，可谓全身是宝，用处多多。

[句子]：苦棟可以治疗头癬

[规则]：无。

[问题]：1. 对[文章]与[句子]相关的内容进行总结。2. 分析文章与句子的相关性。3. 判断[文章]是否能推出[句子]。

回答：1. 总结。[文章]提到苦棟子对于皮肤病、头癬有一定的抗菌止痒的效果。[句子]提到苦棟可以治疗头癬。2. 分析。从[文章]来看，苦棟子可以治疗头癬。3. 判断。是，[文章]可以推出[句子]。

根据给定[文章]、[句子]和一些补充[规则]回答下列[问题]。

[文章]：于人体，可以杀灭水蛭，蚯蚓，蛔虫……苦棟子的出油率很高，因此也用来榨油，制作肥皂，油漆等。冬季寒凉，皮肤易生冻疮，因为油脂丰富，老一辈们也会去采摘一些苦棟子，去核留出皮肉，把它做成油膏，在冻疮上面一抹，涂抹几次冻疮即可痊愈。苦棟子有清热燥湿，行气止痛的功效，煮水可用于胃痛、肝气不舒所致的胸胁疼痛。由于其煎液还有较强的杀菌作用，因此对于皮肤病、头癬也有一定的抗菌止痒的效果。苦棟树还有一个用处，那就是制作装饰品，将其串成手串，拿在手里盘玩，极具特色，曾经也是很多文玩朋友的心爱之物。“棟花香，棟果苦”。苦棟树的花，叶，果，皮，根皆可入药，可谓全身是宝，用处多多。

[句子]：苦棟可以治疗头癬

[规则]：1. 有效果不代表可以治疗。

[问题]：1. 对[文章]与[句子]相关的内容进行总结。2. 分析文章与句子的相关性。3. 判断[文章]是否能推出[句子]。

回答：1. 总结。[文章]提到苦棟子对于皮肤病、头癬有一定的抗菌止痒效果。[句子]提到苦棟可以治疗头癬。2. 分析。抗菌止痒效果与治疗头癬并不完全等同，抗菌止痒效果只是在一定程度上减轻或抑制头癬的发生，而不能完全治疗头癬。3. 判断。否，[文章]不能推出[句子]。

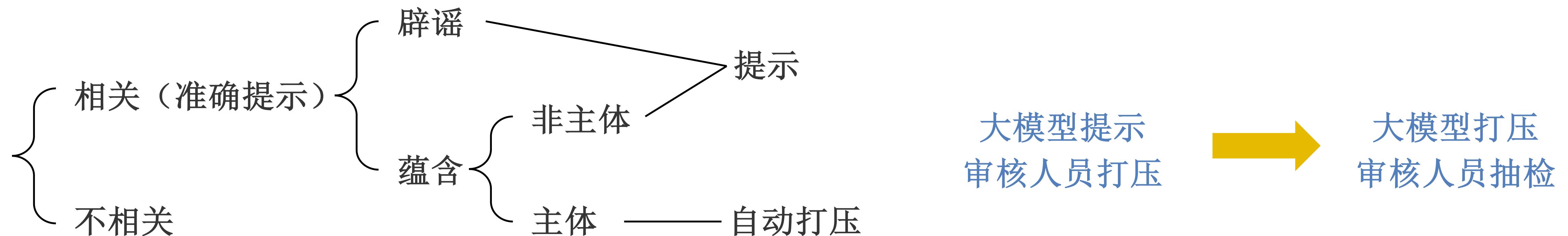
通用审核模型 — 未来工作计划

模型基础能力提升

程度问题 链式包含问题

1. 在固定现有模型的情况下，对于每周做错的案例进行规则补充，引导模型正确推理
2. 补充规则不能解决所有问题，还需要在推理过程中添加对规则的使用，收集一定量做错数据后，将其加入训练集重新训练模型，提高其分析和使用规则的能力

在知识库上实现自动打压



[文章]: 柚子当中含有大量的维生素C, 维C不仅有美白的作用, 还能抵抗坏血块, 还能增强免疫系统。

[句子]: 柚子可美白