

项目一：93 阅兵 & 抗战

项目一：93 阅兵活动 & 抗战历史话题资源识别能力建设

项目背景：面向“抗战胜利 80 周年”重大舆情节点的安全防控需求，针对“阅兵活动”“抗战历史”等高敏话题在内容识别和召回上的覆盖不足问题，构建专项识别能力与召回信号体系，为安全初审与复审环节提供高精度分流支撑。

核心工作：

- 数据分析与样本构建：**围绕“历史阅兵”“93 阅兵”等主题，构造正负样本并抽取关键要素（标题、过检样本、代表人物如“普京出席”等），沉淀主题理解标准与语义特征。
- Prompt 工程优化：**基于 DeepSeek-V3 大模型，设计分层 Prompt 模板体系；通过 Few-shot 语料增强与任务分层设计，实现短平快类召回任务的泛化与高精度识别。
- 效果迭代与鲁棒性提升：**针对误召与漏召样本进行系统分析，引入 Few-Shot 样本补全与对抗样本增强，有效提升模型对模糊语义与影射性表达的鲁棒性。

技术难点与解决方案：

- 难点：**阅兵类与抗战历史类语义重叠度高，对抗样本多、隐式表达频繁，传统规则模型难以覆盖。
- 方案：**结合 Prompt Versioning (v3.1→v3.3) 策略与 Few-shot 伪思维链条增强，在保持高精度的同时提升召回广度与稳定性。

项目成果：

- 阅兵活动相关资源召回率提升至 **99.83%**，抗战历史话题召回率达 **98.2%**。
- 图文、短视频、动态、小视频多业务线综合检出率约为 **0.04%**。
- 模型服务已封装上线，融入安全初审与复审全链路，为后续大模型语料建设提供高质量召回数据支撑。

20w * 800

一、背景与问题

基于抗战胜利80周年的管控背景，需要针对谈论【抗战历史 和阅兵活动 2个话题】相关的资源，或者评论本身，建立专项识别能力，圈定后再继续进行后期模型的处理。针对“抗战胜利纪念日（9·3 阅兵）与相关抗战历史内容”进行模型召回。

在初版 prompt 下，模型存在以下问题：

- 早期 prompt (V1、V2) 虽然能理解语义，却**过于保守**，导致以下问题：
 - 漏召严重：**
模型会认为部分带“阅兵”但偏向外交/人物新闻的内容不是“重大阅兵”；
例如标题：“特朗普访华日期曝光，李在明或错过阅兵”被误判为“外交新闻”。
 - 模型语义歧义容忍度太高：**
当标题含有模糊比喻（如“阅兵是面子，问题是里子”），模型会误判为“比喻用法”而不召回。
 - 召回目标偏向‘精确率’：**
由于前两版 prompt 使用“少量示例 + 条件限定”，模型整体更倾向谨慎输出，造成 Recall 偏低（约 88%）。
- 召回率偏低（约 82%）：**大量真实的抗战/阅兵文本未被识别（如“盛大的阅兵仪式”、“国旗护卫队”被误判为一般庆典，也就是普通阅兵和93 专项阅兵会混淆）。
- 误召问题严重：**模型误将“阅兵式美妆”“校园阅兵训练”等非政治类内容召回。
- 短文本与模糊表达识别弱：**UGC/短视频标题类文本表达简略（如“93来了”“致敬先辈”），模型缺乏上下文推理能力。

5. 输出标签是 ['抗战历史']、['重大阅兵']

二、数据

阅兵活动相关资源召回率提升至 99.83%，抗战历史话题召回率达 98.2%；图文、短视频、动态、小视频多业务线综合检出率约为 0.04%

三、迭代动机与核心目标

为了解决上述问题，Prompt 工程的目标是：

- 提高抗战历史 和 阅兵活动 核心事件的识别召回率（Recall）
- 降低误召与模糊表达的干扰
- 让模型理解“93阅兵”语义的事件级上下文，而非关键词匹配

三、Prompt 迭代历程与关键修

阶段	Prompt 核心策略	存在问题	迭代思路	效果提升
V1 初版 Prompt	直接以关键词触发，如“阅兵”、“抗战”、“9月3日”、“胜利纪念”等	模型高依赖关键词，无法识别语义类表达，如“天安门广场仪式”	引入事件描述式 Prompt，将任务目标改写为“识别所有涉及国家级抗战纪念与阅兵仪式的文本”	召回率由 82% → 88%
V2 优化版 Prompt	增加 Few-shot 示例（正例：93阅兵活动、军队方阵描写；反例：学校军训、节日庆典）	模型在模糊语境下仍会混淆“纪念活动”与“阅兵”	增加对比示例说明（即说明哪些是非阅兵类庆典），并强化“国家级事件”概念	召回率提升至 91.3%

第三版提示词采用了“语义强约束 + 全覆盖策略”，逻辑是：

“宁可误召，也不漏召；只要文本涉及抗战/阅兵相关关键词，就必须标注。”

核心设计思想如下👉：

模块	设计内容	设计目的
「判定标准」部分	明确告知模型“出现关键词即触发标注”，无论上下文是否比喻、娱乐化、延伸语境	打破模型的“语义保守性”，确保召回全面
「负面样例约束」部分	反向定义：即使文本主题偏移（外交、影视、评论），只要含关键语义仍必须标注	避免模型误以为“语境不纯”就不召回
「输出要求」部分	以 JSON 格式输出，强制模型结构化回答并附理由	提升模型输出稳定性，便于批量解析与统计
关键词归类策略	将阅兵与抗战关键词分层管理（阅兵类、战争类、人物类、事件类）	方便模型进行多

四、优化效果与结果

- 最终召回率：从最初 82% → 99.83% 、 98.2%
- 实际业务效果：93阅兵、抗战纪念类内容在全量安全初审与复审召回中覆盖率明显提升，有效降低了人审漏召成本

项目成果：

- 阅兵活动相关资源召回率提升至 **99.83%**，抗战历史话题召回率达 **98.2%**。
- 图文、短视频、动态、小视频多业务线综合检出率约为 **0.04%**。
- 模型服务已封装上线，融入安全初审与复审全链路，为后续大模型语料建设提供高质量召回数据支撑。

项目二：多模态审核模型优化——涉政领域

• **项目二：多模态审核模型优化——动态场景涉政风险识别能力增强**

项目背景：在动态业务场景下，涉政治及领导人相关内容存在明显漏检问题（运营 PM 通报典型 Case）。原多模态审核模型（Qwen2.5-VL-7B 基座）主要基于图文语料训练，导致动态分线召回不足。为此，针对涉政风险四大类（高危反动、涉 X 类、时政有害、涉政风险）进行专项优化，构建动态业务线专属检测能力。

核心工作：

1. **数据分析与模板构建：**结合 PM 通报样本与风险事件数据，分析漏检分布并归纳高危类型模板，并重新统计相关数据配比和量级。
2. **模型结构优化：**在 Qwen2.5-VL-7B 基座上进行多 LoRA 结构优化——涉政治组（领导人脸、涉政文本风险）与敏感场景组（色情、旗帜、暴力等）独立参数更新，减少任务干扰。
3. **领域自适应训练：**基于 LLaMA-Factory 框架进行有监督微调，引入滑动窗口与重叠切词机制，增强模型对长文本混合输入的上下文理解与语义关联捕捉。
4. **服务部署与效果验证：**使用 vLLM 构建高并发推理服务，在专项高危样本集与随机样本集上进行多轮效果验证与压力测试，输出多标签识别结果（leader / text_xi / text_policy / normal），确保服务稳定性与预测准确性。

难点与问题解决：

1. **动态内容语义模糊：**动态场景中短文本上下文信息稀疏，语义表达隐晦，传统模型难以捕捉；**解决：**通过滑动窗口切分与上下文重叠训练，增强模型对片段化语义的连贯理解。
2. **多任务训练冲突：**涉政内容与敏感场景任务差异大，联合训练易导致性能下降。**解决：**采用分组 LoRA 结构进行参数隔离，实现多任务协同优化与稳定收敛。

项目成果：

1. **从 0→1 跑通闭环：**“Excel 标注 → 违规段抽取 → JSON 样本 → LLaMA-Factory 训练 → vLLM 验

证”全流程打通。

2. **数据与配置可复用：**完成四表到 SFT 的自动化转换脚本与数据集注册，沉淀 Qwen2.5-VL-7B + LoRA（rank=16）的稳定训练模板。
3. **为后续扩展预留接口：**为后续测试更大参数量级模型效果，将 qwen2.5-vl-32b 模型作为基座，提供经验。

一、项目背景与动机

在内容安全多模态审核系统中，我们发现**动态业务场景**（如短视频标题、评论、直播弹幕）的涉政、领导人识别存在明显漏检。

- 运营 PM 多次通报：领导人相关、涉政风险内容漏出，尤其在短文本、隐晦表达中；
- 根因分析：动态发文简短，与评论相近，含有很多映射。和图文数据的长篇大论、直白不同。而训练之前多模态大模型审核员收集的都是图文的数据。

~~原模型（Qwen2.5-VL-7B 基座）主要在图文样本上训练，语料偏“长篇直白”，而动态发文多为短句 + 模糊上下文，导致模型召回不足。~~

✓ 简单一句话总结：

原模型擅长识别“图文长内容”，但对“短文本 + 隐晦语义”不敏感，导致动态线漏召严重。

二、优化目标

不明显增加人审压力的情况下，减少涉政涉习漏出：【在尽量保证原来判白率80%左右，不大幅下降（78% 76%）的情况下，召回率提高98%以上】

- 涉政、涉习类风险召回率显著提升（目标 > 98%）；
- 判白率（误召率）基本保持稳定（从 80% 稳定至 76%-78% 区间）。

三、核心技术方案

(1) 风险样本分析与模板构建

- 收集运营 PM 通报的典型漏出 Case；
- 对样本进行分类：高危反动 / 涉X类 / 时政有害 / 涉政风险；
- 针对每类风险构建模板（如leader、text_xi、text_policy、normal）等），用于数据扩充与 few-shot 对比提示词生成。

(2) 多 LoRA 分组优化

- 基座模型：Qwen2.5-VL-7B
- 结构优化逻辑：

原本的涉政模型（了解）qwen2.5 VL 7B模型 在风控垂域模型 覆盖的风险类型【领导人脸、涉政风险相关（细分高危反动）、涉xi（纯文本）、色情（纯图像）、敏感场景（纯图像 64 旗帜 标识 血腥暴力）等

现在 将模型拆为两个 LoRA 子组，分别独立训练：调整成多lora一个基座模型，风险类型分为两个组别

- 涉政治组：领导人脸、涉政风险相关（细分为高危反动）、涉xi（纯文本）；
 - 敏感场景组：色情（纯图像）、敏感场景（纯图像）、64、旗帜、标识、血腥暴力等。
- 优点：任务解耦 → 减少多任务干扰 → 各组参数独立收敛。

(3) 领域自适应训练（LLaMA-Factory 框架）

- 为解决短文本语义缺失问题，采用：
 - 滑动窗口 + 重叠切词机制：在训练中保留上下文衔接； cutoff_len: 8192 overlap: 200
 - 对长短混合文本进行局部上下文增强；
- 训练监控指标包括 loss 曲线、召回样本拟合情况；
- 对高危样本倾向性过拟合，使模型更敏感于潜在风险信号。

(4) 服务部署与验证（vLLM）

- 模型微调后通过 vLLM 部署推理服务；
- 输出多标签结果： leader / text_xi / text_policy / normal；

- 验证集：
 - 专项召回集（PM 通报样本）
 - 随机样本集（代表日常内容流）
- 测试指标：召回率、判白率、漏召率、推理延迟。

四、核心难点与解决方案

难点	描述	解决方案
动态文本语义模糊	动态短文本上下文稀疏，隐喻多，难识别	滑动窗口 + 重叠切词机制，增强上下文连贯性
多任务训练冲突	涉政与色情类任务差异大，联合训练不稳定	分组 LoRA 参数隔离，独立优化两类任务
召回与判白权衡	强召回导致误召上升	在提示词和标签判定逻辑中加入置信度阈值约束，控制误召上限
高危样本偏分布	数据极度不平衡，高危样本稀少	通过模板扩充 + 难样本复采样，增强训练覆盖度

五、效果与结果总结

指标	优化前	优化后	改进幅度
动态涉政召回率	约 80%	98%+	↑ +18%
判白率（不涉政误召）	约 80%	76%-78%	≈ 持平
漏召率	约 20%	<5%	↓ -15%
推理性能（vLLM）	稳定支持高并发	✅ 通过线上压测验证	

项目三：3S 词表 & 业务线词表梳理与调整

• 项目三：3S 词表 & 业务线词表梳理与调整

项目背景：PGC 内容送审中 3S 词表触发量约 2.5%、业务线词表约 1%，规则“直送”导致无效送审比例偏高。为降低送审量级、避免风险漏出，开展 3S/业务线词表清洗与分级改造，目标将整体送审率压降至 2% 以下。

核心工作：

1. 数据抽取与物料清洗：在 007 机器上使用自研脚本从物料库批量拉取图文业务线样本，包括 7.28 当天 3S 高危词送审数据及近 7 天“3S 词表送审且人审拒绝”样本，提取对应的标题、正文、图片链接、OCR 内容；完善 Excel 取词正则与 NID 精度处理，确保词表与物料一一对应。
2. 模型部署与逻辑改造：离线部署“词表分级大模型”，修改送审逻辑，调整请求参数模板——高危词命中样本不再直接送审，而需继续过词表分级大模型进行复核。
3. 风险召回与豁免测试：分两路测试词表分级大模型对于 3s 词表命中数据的风险召回及豁免能力：
 - A) 豁免测试：拉取 7.28 一天 PGC 业务线 3S 词表高危词命中数据；豁免率 78.56% (2092/2663)
 - B) 召回测试：拉取近 7 天“3S 词表送审且人审拒绝”样本；风险召回率 34.87% (902/2587)
4. 结果回收与策略反馈：将评测结果回传 PM/运营复核，协助判定高危词降级与保留策略，形成可量化的“高危 → 低危”评估闭环。

项目成果：

1. 跑通词表清洗 → 词表分级大模型 → 结果评测 → 策略反馈全链路流程。
2. 验证词表分级模型在涉政领域的可行性，为后续“图动短小”等多业务线词表治理提供标准化清洗与评测模板；输出可复用脚本与参数模板，为后续词表精细化治理奠定基础。

一、项目背景

在内容安全体系中，3S 词表 (Sensitive/Security/Society) 与业务线词表是文本审核中最基础的一环。但在 PGC 内容送审链路中出现了明显问题：

- 3S 词表触发率约 2.5%，业务线词表约 1%，
- 其中大部分内容其实是“低风险命中”，被 规则直送人审，导致：
 - 无效送审量高；
 - 审核队列压力大；
 - 真正高危样本比例被稀释。

因此，我们发起了“3S / 业务线词表清洗与分级改造”专项，

在不增加漏召风险的前提下，将整体送审率压降至 2% 以下。

二、核心问题分析

问题点	影响	解决思路
1 高危词规则过严	误触大量低风险文本	引入“词表分级模型”做语义层复核
2 词表与物料脱节	同一个词在不同语境下风险不同	拉取真实物料（标题+正文+OCR）重建语义上下文
3 缺乏量化指标	难以评估词表降级安全性	设计“豁免率 / 召回率”双指标评测体系

三、优化方案与实施路径

(1) 数据抽取与物料清洗

- 在 007 机器 上使用自研脚本批量拉取图文业务线样本：
 - 7.28 当天 3S 高危词命中送审数据
 - 近 7 天 “3S 词表送审且人审拒绝” 样本
- 对每条物料提取：
 - 标题 / 正文 / 图片链接 / OCR 内容
- 用 Excel 正则和 NID 处理逻辑对齐，确保 “词表词项 ↔ 样本语义” 精确对应。

✅ 这一步确保了模型输入数据和规则词项的“语义一致性”，避免虚假命中。

(2) 模型部署与逻辑改造

- 离线部署“词表分级大模型”（基于轻量化语义模型）；
- 改造原有送审逻辑：
 - 高危词命中样本 → 不再直接送审 → 改为经大模型复核。
- 模型输出高危/低危标签及风险理由；
- 同时改写审核请求模板以适配模型新输出。

✅ 这一环节的核心是：将原本“规则直送”链路改为“规则+模型联合决策”链路。

(3) 风险召回与豁免测试

对模型的可行性进行 A/B 两路评测：

测试类型	数据来源	评估目的	结果
A) 豁免测试	7.28 当天 PGC 3S 高危词命中数据	验证模型的“降级”能力	豁免率 78.56% (2092 / 2663)
B) 召回测试	近 7 天“3S 词表送审且人审拒绝”数据	验证模型的“风险识别”能力	风险召回率 34.87% (902 / 2587)

✅ 结果表明模型能显著过滤无效送审，同时保持核心风险召回。

(4) 结果回收与策略反馈

- 将模型评测结果反馈 PM / 运营团队；
- 联合判定哪些高危词可降级、哪些需保留；
- 最终形成了可量化的：

- 高危 → 低危降级策略；
- 可视化评测报告 + 模型脚本模板；
- 为后续其他业务线（如图动短小）提供了可复用标准化治理流程。

四、项目成果与价值

维度	成果
流程闭环	跑通“词表清洗 → 模型分级 → 结果验证 → 策略反馈”全流程
指标结果	无明显漏召风险下，整体送审率压降至 < 2%
方法论沉淀	输出词表治理标准化模板与脚本，为多业务线复用
实效意义	降低人审压力、提升模型召回精度、构建自动化治理闭环

项目四：

Token储备量统计：

	数据类型	Token数（单位：M）	是否需要清洗 /二次清洗	备注
1				
2	敏感场景、色情、人脸等蒸馏数据	约25M	否	GLM、Ovis、Ernie 已完成
3	黑库数据	约20w图片		待标注
4	wiki数据	约100M	是	待过词表及清洗
5	开源数据MCP-CC	约800B，1.6TB	是	百科数据清洗完成，paper、爬虫、数据数据正在进行中。
6	开源数据 BLIP3-KALE	2.18亿对图像-文本对	否	待转换语言。
7	审查标准	约0.2M	否	目前文字部分已经整理完成；图片部分进展50%；
8	习变体			待爬取
9	敏感人物库		否	已完成
10				

• 项目四：风控多模态大模型预训练数据清洗与去重流水线构建

项目背景：面向 Chinese-Tiny-LLM 预训练任务，负责搭建高质量中文语料预处理与去重流水线，为风控 MLLM 基座增量预训练提供干净、高纯度的数据基础。数据源覆盖百科、书籍、学术论文、Common Crawl 及问答文本五大类，总规模超 1.6TB，约 800B Token。

核心工作：

1. **过滤模块优化：**深入分析 `filter.py` 主脚本，理解其多层次过滤逻辑（格式统一、URL 过滤、自定义规则、语言识别、重复检测），针对百万级输入样本调优 `fastText` 语言识别参数与自定义规则。
2. **并行化数据清洗：**使用 8 个 `workers` 并行运行过滤流程，处理多源数据集（Baike、Books、CC、Papers、Others），实时监控内存与日志状态，完善输出体积监控、日志打印与进度可视化。
3. **去重流水线构建：**实现三层去重机制：全文去重（Bloom Filter）、MinHash LSH 去重、相似行去重，成功跑通 MinHash 生成 → 重复对 → 连通组件 → 重复行 ID → 最终去重全链路。
4. **数据合并与校验：**完成 CC (650GB+) 数据的分片合并与异常修复，解决前导 0 报错及非 `part` 文件合并问题；制定标准化目录结构与日志规范，确保处理结果可追溯。
5. **性能优化与自动化：**优化并发策略（Others/Papers 降至 2-4 `workers`），避免 I/O 竞争；增加行速、处理时间、输出速率等日志指标，构建可复用的高并发数据清洗框架。

难点与问题解决：

1. **难点：**千万级中文语料的并行过滤与去重在内存与磁盘 I/O 上极易瓶颈。
2. **解决方案：**通过任务分批执行、日志级监控与断点重启机制确保稳定运行；对 MinHash 阈值与相似行检测算法进行调优，平衡精度与效率。

项目成果：

1. 跑通从过滤 → 去重 → 合并 → 验证，在 baike (6GB) 数据上的完整预处理链路。
2. 输出符合预训练标准的中文高质量数据集，为后续 Qwen3-VL 风控基座增量预训练提供核心语料支撑。

一、项目背景与目标

当时团队在做 Chinese-Tiny-LLM 的预训练任务，需要为后续的风控多模态大模型（MLLM）提供干净、高质量的中文语料。

但是：

- 原始数据源极其庞大（百科、书籍、论文、Common Crawl、问答，共约 1.6TB，800B Token）；
- 质量参差不齐（重复多、乱码多、网页噪声多）；
- 没有一套完整的、可复用的清洗 + 去重流水线。

于是，我负责从零搭建一套中文语料预处理与去重流水线，目标是为风控大模型的增量预训练构建高纯度、可追溯的数据基底。

✅ 关键词总结一句话：


“我做的是一套能稳定处理 TB 级中文语料的过滤去重流水线，为大模型预训练提供干净的数据输入。”

二、痛点与问题分析


问题	描述	后果
1. 多源数据格式差异大	不同来源（百科/书籍/网页/问答）的结构与字段不一致	无法统一清洗逻辑，脚本容易中断
2. 重复比例高	Common Crawl 与问答类数据存在大量相似内容	影响模型有效 Token 利用率
3. 并行处理压力大	处理千万级样本时 CPU/I/O 占用高、容易 OOM	程序不稳定、进度不透明

三、核心技术方案与实现路径

(1) 过滤模块优化

- 深入分析并重构 `filter.py` 主脚本逻辑，包括：
 - 统一格式校验；
 - URL 过滤；
 - fastText 语言识别；
 - 自定义风险规则过滤；
 - 重复检测模块。
 - 针对百万级样本，调优 fastText 参数（`-minCount`, `threshold`）以提升中文检测准确率。
 - 增加日志与错误处理逻辑，让过滤可断点恢复。
-  解决痛点：清洗逻辑碎片化 → 模块化与容错化。

(2) 并行化数据清洗

- 采用 **8 worker 并行过滤框架**，对五大类语料（baike/books/cc/papers/others）分任务执行；
 - 通过 `ionice + nice` 控制 CPU/I/O 优先级；
 - 实现日志心跳与处理速率监控（行速、输出速率、文件体积）；
 - 输出日志以分钟粒度统计进度与处理吞吐。
-  解决痛点：百万级样本耗时长、进度不可控 → 高并发 + 可视化进度追踪。

(3) 三层去重机制构建

设计并跑通完整的 三层去重链路：

层级	方法	目标
① 全文去重	Bloom Filter	过滤完全相同文档
② 相似度去重	MinHash + LSH	检测相似文本块（语义重复）
③ 行级去重	行哈希比对	去除相似句或重复段落

并完成从
MinHash 生成 → 重复对构建 → 连通组件聚合 → 重复行 ID 映射 → 去重结果输出
的全链路打通。

🎯 解决痛点：重复内容消耗 Token → 精准去重、节省算力。

(4) 数据合并与校验

- 处理 CC 数据（650GB+）分片合并问题；
- 修复 `part` 文件合并异常与前导 0 报错；
- 定义统一目录结构与命名规范；
- 增加日志验证脚本，确保每一阶段可追溯。

🎯 解决痛点：TB 级数据分片管理复杂 → 结构化目录+日志校验。

(5) 性能优化与自动化

- 根据数据源差异动态调整并发策略：
 - Papers/Others 降为 2-4 workers，避免磁盘竞争；
- 增加指标：行处理速率 / 输出速率 / 耗时统计；
- 构建一套 可复用的高并发数据清洗框架。

🎯 解决痛点：系统不稳定 → 自动化与监控化。

四、难点与解决方案

难点	具体问题	解决策略
内存与 I/O 瓶颈	过滤与去重同时运行时容易卡死	分批执行 + I/O 优先级控制 + 断点重启机制
MinHash 调参难	阈值过高漏检，过低误删	动态调优阈值（0.75-0.85）并人工抽检校正
大文件合并报错	非 <code>part</code> 文件参与合并、文件头异常	增加正则过滤与文件头检查脚本

五、结果与价值

成果	描述
✅ 完整链路跑通	实现从过滤 → 去重 → 合并 → 验证的可复用流水线
✅ 高质量语料产出	清洗后的语料符合预训练标准，噪声率显著下降
✅ 工程价值	输出高并发处理模板和日志体系，为其他语料预训练任务复用
✅ 实际支撑	支撑后续 Qwen3-VL 风控基座 增量预训练任务的核心语料构建