蒋炳杰

资深后端/AI工程师 | 技术负责人

- 基础信息:男/1993/未婚/在职/中国浙江省温州市/有护照
- 学历：本科/中国人民大学/计算机科学与技术/2018年
- 目前所在城市：北京
- 工作年限：9年
- 到岗时间：一个月以内
- 手机：+8617681718233（联系时间：10：00-19：00）
- Email：jiangbingjie1218@gmail.com
- QQ/微信号：284474102/a67793581
- GitHub：http://github.com/a67793581

专业摘要

- **资深后端开发工程师与技术负责人**，拥有 **9+年** 互联网行业经验，专注于**高并发系统架构设计**与**AI应用工程化落地**。擅长带领团队从0到1打造产品，并对技术选型、性能优化与成本控制有成功实践经验。
- **AI工程专家**：深耕LLM、AIGC、Agent等前沿领域，具备丰富的**模型微调（TTS, 多模态）、AI Agent中台构建、向量数据库应用与模型部署优化**经验，能将算法能力转化为稳定、高效的生产服务。
- **高并发架构师**：具备亿级用户流量产品的设计与优化经验，精通**微服务治理、分布式缓存、数据库分库分表、消息队列**及复杂系统（如IM、直播）的平滑演进与稳定性保障。
- **开源贡献者**：积极参与技术社区，是多个知名开源项目（如menGPT, go-mcp）的代码贡献者，具备良好的技术视野和工程规范意识。

工作履历

万格创新(北京)智能科技有限公司（Sapient Intelligence） | AI Engineer / FDE
2025年12月 - 至今 | 官网：https://sapient.inc/

负责 Agent 产品的架构、模块边界和上线收口。

- **ImmiClaw** | 技术负责人
  - 已上线的移民文书 Agent。项目地址：https://immiclaw.cn/
  - 技术选型：Python、FastAPI、LangChain、LangGraph、SSE、PostgreSQL、Redis、Milvus；浏览器工具用 Chrome Extension 和 CDP。
  - **多 Agent**：一个主 Agent 包办检索和四类文书时，工具和上下文混在一起，单轮塞不下。主 Agent 只负责调度和综合评估；research、简历、推荐信、个人陈述、合作意向各自编译成独立 Agent，自带 prompt、工具、middleware 和 checkpoint。主 Agent 用 `task` 下发任务说明和类型，用 `session_id` 续跑同一段；子 Agent 只带回结果，不把过程写进主上下文。
  - **工具和技能**：子 Agent 不继承主 Agent 工具。检索只有搜索和截图，写作才有 DOCX。技能是只读目录里的 `SKILL.md`，按案件类型挂给对应角色；摘要随调用注入，细则由 Agent 自己读文件，避免整包塞进 prompt。
  - **长任务续跑**：检索和文书要跑几分钟，SSE 一断或用户一刷新，进度就丢。子会话状态落在 checkpoint 和 Redis，中断后按 `session_id` 从断点接着跑，同一会话加锁避免并发重入。

    - **RAG**：判例在官网，对话和案件各建索引会给出不同依据。爬取 EOIR/AAO，抽取后双写 PostgreSQL 和 Milvus，日更和存量走同一套入库。
    - **浏览器工具**：登录、验证码和国内不可达的站点，模型自己点不完。tool call 才下发打开、点击、输入、截图和 PDF；卡住等人，完成后自动续跑。截图用浏览器池，打不开的页面转到日本节点。
    - **模型与权益**：六家模型接口不一致，试用和正式订阅如果各算各的，创建、评估和终审会扣错。统一模型工厂收口调用；案件三个阶段共用同一套权益校验。

- **行测出题 Agent**
    - 内部出题工具。按题型生成行测题，覆盖言语理解、言语表达、资料分析和判断推理，生成后必须去重和质检。
    - 技术选型：Python、Code Skills、Milvus、Go CLI。负责工作流和技能，从第一版收到可重复跑的质检门禁。
    - 言语理解、言语表达、判断推理各做一个 skill；Milvus 做向量去重；难度、子题型和通过标准统一。先去重再质检，不通过停在当前题，不能进入下一题。

- **大模型 API 代理**
    - 内部统一入口。按端口转发 OpenAI、Anthropic、Gemini、xAI，去掉客户端 IP 和来源头，加上 TLS 和日志轮转。技术选型：Go、Docker。独立开发。
    - 业务侧不再各自持有厂商地址和密钥，调用、排查和轮转日志都走这一层。

北京一点网聚信息技术有限公司(北京弥特信息技术有限公司) (METADREAM) | 资深后端/AI工程师 -> 技术负责人
2023年6月 - 2025年12月

负责公司创新AI业务线的技术工作，主导多个从0到1的AI产品研发，作为技术负责人攻克多项技术难关，成功将AI技术应用于生产环境。

- **LaterCast（AI播客小程序）| 技术负责人 & 唯一后端**
    - **主导项目**：负责全部后端与AI模块架构设计及开发。
    - **AI工程化**：基于PyTorch与indexTTS模型，搭建了**全自动化说话人克隆流水线**（音频下载->相似度分类->降噪->微调训练），将新声音模型的上线周期从周级缩短至天级。通过VLLM推理引擎与Docker容器化部署，保障服务稳定性。
    - **效果**：构建的AI语音MOS评分达4.2分，为产品提供了核心差异化能力。

- **Bondee Pal / Blopee（AI社交应用）| AI中台架构师**
    - **架构设计**：主导设计并实现了**AI Agent中台**，提供可视化工具编排与多Agent协作工作流（基于ReAct模式的意图Agent与专家Agent协作）。
    - **核心技术**：基于MemGPT思想，设计并**使用Java自研了高性能无限记忆系统**，采用“DB+向量数据库”双写存储架构，成功将记忆读写QPS从10提升至100，延迟降至100ms以内，性能提升10倍，完美支撑了海量用户与AI的长期个性化互动。
    - **工程落地**：通过该中台，极大降低了业务方接入AI能力的门槛，新Agent功能开发效率提升80%。

- **ZapArt（马良）| 项目负责人**
    - **从0到1**：独立设计并开发基于ComfyUI的动态绘图工作流服务，为公司内部AIGC内容生产提供了强大且灵活的自动化工具，替代了外部采购方案。

- **Bondee / 啫喱（亿级社交应用）| 核心贡献者**
    - **高并发架构**：主导了**群组系统“读扩散”重构与分库分表**项目，支撑了**3万人超级大群**场景，使系统容量实现无上限水平扩展，消息接口耗时降低60%以上。

    - **AI赋能业务**：基于QWen2-VL-7B模型微调训练**AI审核模型**，准确率94%，使人审占比下降30%，每年节省大量人力成本。
    - **成本优化**：通过设计通用接口层，主导完成直播CDN从自研到多第三方厂商的接入，**降低带宽成本30%**。

北京密境和风科技有限公司（花椒直播） | 高级后端开发工程师
2021年12月 - 2023年5月

- **直播系统优化**：负责直播、连麦、IM系统核心功能开发与维护。设计通用CDN接口层，解决了多厂商接入差异问题，提升了系统的灵活性与可维护性。
- **IM系统上云**：主导IM系统微服务化改造与**优雅下线方案**设计（通过服务注销、排水、等待超时断连等机制），实现了有状态长连接服务的平滑滚动发布，保障了上云过程用户无感知。

北京蜜莱坞网络科技有限公司（映客） | 后端开发工程师
2021年5月 - 2021年10月

- **对缘（视频相亲平台）| 核心系统开发者**
    - **负责系统**：主导好友、动态、活动及推荐系统的开发、优化与迭代，支撑日均百万级活跃用户。
    - **技术攻坚**：针对公司自研微服务框架缺乏SDK生成能力导致的开发低效问题，**抽象并设计了一套通用SDK**，统一了微服务间的调用规范，将接口联调开发效率提升约40%。
    - **业务贡献**：所负责的系统为App核心功能模块，直接贡献了公司主要营收来源，通过持续的性能优化和功能创新，提升了用户粘性与付费转化。

杭州空岛科技有限公司 | 后端开发工程师
2019年4月 - 2021年4月

- **商家岛（微信营销SaaS平台）| 后端技术负责人**

    - **项目职责**：作为项目后端主力，**独立统筹设计并开发了超20个核心公共模块**，包括：红包、提现、弹幕、分销、抽奖、客服、资金流水、活动审核等系统，支持30万商家、覆盖上亿消费者。
    - **架构设计**：针对营销活动多、通用性强的特点，创新性地采用**多态表设计**对功能进行抽象，极大减少了重复开发，实现了高可复用性的业务中台架构。
    - **技术栈升级**：推动并实践了从LNMP单体架构到基于**Docker、K8S、Gin的GoLang微服务**架构的演进，为后续系统拆分和高并发场景打下基础。

- **微服务全资源审核系统 | 全栈开发者**

    - **从0到1建设**：独立设计并开发了支持文字、图片、视频的全资源审核微服务，提供同步/异步、批量审核能力，并集成熔断降级机制保障高可用。
    - **技术落地**：深入应用**Kubernetes、Docker、Gin**等技术，成功将服务部署生产环境，日均处理审核任务超百万次，有效拦截违规内容，为公司平台安全保驾护航。

杭州洞见科技有限公司 | 后端开发工程师
2018年10月 - 2019年4月

- **洞见云运营管理系统 & 嗨皮时光APP | 项目核心开发者**
    - **项目职责**：独立负责浙江地区上百家线下门店、数千台娃娃机的云端运营管理平台和移动端APP的后端研发，涵盖权限、采购、财务、统计分析和IoT设备监控等模块。
    - **性能优化**：针对海量运营数据导致的统计报表查询缓慢（超10秒）问题，通过**设计数据预处理脚本与定时任务**，提前聚合关键指标，将报表加载时间优化至3秒内，提升运营决策效率。
    - **物联网实践**：深入参与娃娃机等IoT设备的通信管理（心跳、命令下发），排查并解决因网络丢包导致的设备控制不稳定问题，积累了物联网项目的一线实战经验。

杭州爱聚科技有限公司 | 后端开发工程师
2017年7月 - 2018年10月

- **短信营销系统 | 后端开发者**
  - **核心开发**：负责数据库设计与核心业务代码开发，构建了支撑全国上万家电商业家的短信营销平台。
  - **解决并发难题**：设计实现了基于**Redis锁（SETNX+GETSET）** 的防重复发送机制，并结合前端JS防抖，彻底解决了浏览器卡顿导致的短信重复发送问题，将故障率降至近乎为0。
  - **业务价值**：该系统帮助运营人员高效创建营销活动，将客户转化率从10%提升至15%，成为公司重要的业务工具。
- **语音呼叫系统 | 项目负责人**
  - **从0到1交付**：**独立负责**项目全流程，包括技术选型、硬件部署（语音网关）、服务端开发（Java/PHP）与上线。
  - **全栈开发与兼容性解决**：通过JS调用Java服务实现呼叫逻辑，并**攻克多浏览器兼容性难题**（Chrome, Firefox, IE9），将故障率从30%降低至1%。
  - **成本控制**：项目成功替代了外部方案，**为公司节省了50%的语音系统采购成本**，并奠定了CRM语音系统的基础。

技能清单

以下是我熟悉和使用过的技能：

- AI框架: TensorFlow / LLaMA-Factory / ComfyUI / LangChain / PyTorch
- AI应用: RAG / Agent / AIGC
- AI模型:Stable Diffusion / LLaMA / QWen-VL / IndexTTS
- 编程语言：PHP / GoLang / CSS / JS / HTML / Java / C / C# / C++ / Lua / Python
- 框架：ThinkPHP / CodeIgniter (CI) / Laravel / Swoole / Composer / Gin / Echo / GRPC / Symfony / GraphQL
- 前端框架：Bootstrap / JQuery / HTML5 / Vue / Svelte / React
- 数据库相关：MySQL / Redis / SQL Server / MongoDB / ES / Memcached / Kafka / Pulsar
- 端口代理相关:Traefik / Nginx / Apache
- 版本管理:文档和自动化部署工具：Svn / Git / Jenkins / Supervisor / Docker / Portainer / K8S
- 服务器：Linux / Windows Server

开源贡献

- mcp-go: 贡献了端点接入兼容等功能，被官方仓库合并。
- OpenKF: 为其Go语言开源客服系统贡献了自动化单元测试代码，提升了项目代码可靠性。
- ComfyUI: 贡献了新的队列接口实现，增强了其扩展性。
- letta-ai（MemGPT） 优化了记忆系统索引,提高了它的性能。

项目成果精选

1. **性能优化**：自研Java版无限记忆系统，性能较原Python方案提升**10倍**（QPS 10->100，延迟<100ms）。
2. **成本控制**：通过AI审核模型降低人审成本，年节省预估超百万元；通过接入第三方CDN降低带宽成本**30%**。
3. **效率提升**：搭建AI Agent中台，将新AI功能的接入周期从**2周缩短至2天**，开发效率提升80%。
4. **系统容量**：完成亿级社交应用群组系统重构，成功支撑**3万人大群**，实现系统容量无上限水平扩展。

致谢

感谢您花时间阅读我的简历，期待能有机会和您共事。