Qwen3-Max-Preview vs. Kimi K2-0905｜调研简报（2025-09-08）

Qwen3-Max-Preview（闭源·预览）

- **定位**：阿里云旗舰预览模型，提供 **≈256K** 上下文，面向长文档处理与企业问答。
- **优势**：
  - **快**：非思考架构 + API 工程优化，响应延迟低。
  - **长上下文**：上下文缓存可降低重复计算成本。
  - **商用能力**：阿里云 Model Studio、OpenRouter 多渠道接入。
- **技术报告**：暂无独立 Max-Preview 报告，可参考 **Qwen3 总技术报告**了解架构背景。

Kimi K2-0905（开源·权重可得）

> 面向「Agentic Intelligence」的超大规模 MoE 基座 + 强代理后训练体系；0905 版在 Agentic Coding 与上下文长度（256K）进一步增强，已开放 Base / Instruct 权重。

模型结构特点

- **超稀疏 MoE 基座**：
  - **总参 ~1T，激活 ~32B，Top-8 路由**；
  - 以 **MLA（Multi-head Latent Attention）** 为注意力骨干，面向长上下文推理的 **FLOPs/吞吐** 做了结构侧优化（相较同类，将注意力头数下调以换取长序列推理效率）。
- **面向 Agent 的工程取向**：
  - 以“工具使用 + 多步规划”为第一性能力目标，在架构/后训练全流程对 **Agentic** 场景做增强；
  - 开源 **Base/Instruct** 权重便于私有化与二次研究。
- **上下文窗口**：
  - 技术报告覆盖至 **128K** 的扩展策略；
  - **K2-0905 Instruct** 将服务侧上下文进一步 **扩至 256K**。

附录｜训练与数据（实现细节）

A. 预训练稳定化与配方

- **QK-Clip 机制**：训练步后读取各头最大 attention logit $S^h_{\max}$，当超阈值 **τ=100** 时，仅对 **MLA 的非共享分量**执行缩放（q_C, k_C 乘 $\sqrt{\gamma}$，q_R 乘 γ，共享 k_R 不动），**不改变当步前后向**；早期常激活，收敛后**自停用**。
- **配方要点**：总计 **15.5T tokens**；主程窗口 **4k**，尾段以 **YaRN** 激活 **32k/128k**；学习率与 batch 采用“预热-持平-余弦衰减”。

B. 数据：提升 token-utility

- **知识/数学重写**：以**分块自回归重写**替代多 epoch 重复，结合多风格/多视角提示，提升**每 token 学习信号**；多语互译与题型改写覆盖知识/数学等域。

C. 后训练：SFT → 联合 RL

- **SFT**：延续 **Muon** 做指令微调；多域高质数据（含模型与人审）过滤。
- **Agentic 数据合成**：真实/合成工具池 + 任务/代理/轨迹三级生成与判别过滤，规模化获得**可验证工具使用轨迹**。
- **统一 RL**：
  - **RLVR**：在可判定任务上以**二值可验证奖励**直接优化“做成事”。
  - **自评 Rubric 奖励**：使用器对生成进行成对排序/打分（偏好、深度、事实性、安全等维），与 RLVR 形成闭环，抑制 reward-hacking。

D. 系统与长上下文

- **并行与内存**：报告描述了 PP/EP/DP 组合与重算、FP8 激活存储、CPU-offload 等手段以控制显存峰值并重叠通信；为 **128K** 长上下文推理打基础。**0905** 在线服务侧将窗口拉升到 **256K**。