

Bottom-Up Token 预测速查表 · 七步法速查 · v1.0 · 2026.09

① 核心公式

T_period = SUM_i [N_users x freq_i x (L_in_i + L_out_i) x amp_i] x ai_share

② 七步法

Step	动作	关键产出
1	业务场景 → 任务清单	10-30 个原子任务 (Task ID)
2	单次任务 → Token 拆解	System + User + Context + Output
3	调用频次估算	按用户分层加权 (高频/中频/低频)
4	基础聚合公式	T_day = SUM (DAU x freq x token)
5	Agent / RAG 放大	amp ∈ [1.0, 10]
6	场景敏感性	P10 / P50 / P90 三档
7	历史校准 + 回归	OLS + MAPE < 20%

③ amp 系数速查

组件	amp	场景
基础对话	1.0x	单轮
多轮对话	1.5-2.5x	历史+重述
RAG 检索	2.0-4.0x	Top-K 文档
Agent 工具	3.0-8.0x	多次 LLM
多模态	1.5-3.0x	图像转 token
综合叠加	4.0-10.0x	RAG + Agent

④ 三场景敏感性调节

维度	P90	P50	P10
DAU (QoQ)	+10%	+30%	+60%
会话长度	持平	+15%	+30%
Agent 渗透	5%	20%	45%
amp 加权	1.5x	2.5x	3.5x
AI 承接率	50%	70%	85%

⑤ 关键阈值 (红线)

指标	红线	动作
MAPE > 20%	回归失效	重检 Step 1-3 假设
差异 > 30%	BU vs TD 不一致	重新审视假设
amp < 1.0	非法值	必填 ≥ 1.0 ; Agent ≥ 2.0
LTV/CAC < 3	不健康	产品 / 定价调整
AI 成本/毛利 > 30%	超警戒线	优化模型或提价

⑥ Excel 模板字段对照

字段	类型	示例	说明
task_id	Str	T01	任务 ID
tokens_in_p50	Int	500	输入 token 中位数
tokens_out_p50	Int	120	输出 token 中位数
dau_segment	Int	5000	分层日活
freq_per_user	Float	3.5	单用户日均频次
amp_factor	Float	2.5	放大系数
ai_share	Float	0.7	AI 承接率
daily_tokens	Formula	=B*C*...	日均自动计算
p10/p50/p90	Float	1.5/2.5/3.5	三档敏感性

⑦ 5 天实施清单

Day	动作	产出
1	三方对齐业务场景	场景清单
2-4	抽样测 Token + 拉 DAU/频次	频次矩阵
5	聚合 + 放大 + 双校验	季度预测