习题解析与实验指导

LLM from Scratch: Zero to Hero

韩威

配套学习用书

使用说明

《习题解析与实验指导》与教材配套使用. 教材承担概念、原理、推导和教学例题，本册集中安排习题、实验指导及参考解析. 两册的分工是先理解机制，再通过推导、实现和证据检验理解.

第一部分按教材五篇 42 章的顺序编排；第二部分按五类主题编排 24 个 Notebook 实验，与教材形成多对多关系；第三部分汇集 504 份习题解析；第四部分列出实验指导实际引用的论文与框架资料. 教学例题及其完整计算过程位于教材.

按方向选择习题及实验

先按教材阅读说明建立共同基础，再选择应用智能体系统、模型训练方法或训练推理系统方向. 习题按对应教材章节选择，实验按下面的主题顺序进入；实验依次按语言模型理论基础、应用智能体系统、模型训练方法、训练推理系统、模型架构拓展编排，各实验的输入资产和知识前置见其学习契约.

共同基础的机制实践由实验 1、实验 2、实验 3、实验 4 和实验 5 依次支撑，覆盖基础复习、数据、分词、Transformer 和自回归生成. 完整实验包含训练与数值验证，阅读教材的共同基础可以按主题选读. 实验 6 提供各方向共用的评估方法，实验 7 提供安全分析；使用第三方开放权重制品时，结合实验 21 核对来源和模型仓库.

应用智能体系统沿实验 6、实验 8 和实验 9 学习评估、提示策略、RAG 与有界智能体，提前掌握实验 7 中的威胁模型和工具授权. Harness 的持久化、隔离与恢复机制继续结合教材智能体架构和运行系统研读；实验 9 的实践范围是检索、工具、工作流与有界执行循环.

模型训练方法在掌握训练循环与监督目标后进入实验 11 的监督微调主题及实验 12，深入时选择实验 10 的预训练、实验 13 的数据工程或实验 11 与实验 14 的人类反馈、偏好对齐和推理训练. 实验 11 将多类目标放在同一文件，主题阅读可分阶段，运行时仍按文件内数据与定义依赖执行.

训练推理系统以实验 15 建立资源预算. 推理服务分支沿实验 6、实验 17、实验 7 和实验 18 学习评估、部署、安全与优化；训练系统分支在完成实验 10 的训练与恢复知识后进入实验 16. 实验 19 作为模型优化专题，涉及蒸馏和训练时量化时补齐训练前置. 多模态与其他模型专题根据任务选择实验 23、实验 20、实验 22 和实验 24.

作答及运行

外侧页边的危险弯道标记表示首次阅读可暂时跳过的扩展题，其他题目不加标记. 建议先独立作答，再核对解析. 实验先提出预期与失败条件，再运行代码并记录实际结果，不能以现有输出替代自己的复现证据.

配套 Notebook 位于项目的 `notebook/` 目录. 指导中的文件名和函数定位依据当前仓库源文件；环境、资源和外部制品要求应与 Notebook 中的依赖及配置单元共同核对. 实验结果以读者实际执行的环境、输入和运行记录为准.

两册使用共同符号与参考文献库. 习题不要求查阅文献；实验指导引用的方法论文与框架资料集中列在第四部分. 跨册公式与图表以教材编号定位；发布时应成套更新两册，避免版本不一致.

目录

使用说明 i

第一部分 习题 **1**

第一章 绪论 2
第二章 数学基础 3
第三章 神经网络基础 4
第四章 文本表示 6
第五章 注意力机制 8
第六章 Transformer 网络结构 9
第七章 位置表示 10
第八章 自回归语言建模 11
第九章 模型评估 12
第十章 模型安全 14
第十一章 上下文学习 16
第十二章 信息检索 18
第十三章 检索增强生成 20
第十四章 领域应用 22
第十五章 智能体架构 23
第十六章 智能体运行系统 25
第十七章 优化及泛化 27
第十八章 训练数据工程基础 28
第十九章 训练稳定性 30
第二十章 监督微调 31
第二十一章 参数高效微调 32
第二十二章 训练工程优化 34
第二十三章 预训练原理 35
第二十四章 强化学习基础 37
第二十五章 人类反馈学习 38
第二十六章 可验证推理训练 39

第二十七章　模型规模及资源预算 41
第二十八章　算力基础设施 42
第二十九章　分布式训练 44
第三十章　增量解码 45
第三十一章　批处理调度 46
第三十二章　模型服务架构 48
第三十三章　推理计算优化 49
第三十四章　模型优化 50
第三十五章　双向语言建模 53
第三十六章　混合专家模型 54
第三十七章　长上下文建模 55
第三十八章　开源模型体系 56
第三十九章　视觉 Transformer 57
第四十章　多模态建模 58
第四十一章　扩散模型的概率原理 60
第四十二章　条件视觉生成 61

第二部分　实验指导 **63**

实验总则 64
实验 1　神经网络训练基础 66
实验 2　训练数据工程 68
实验 3　子词分词 70
实验 4　Transformer 实现 72
实验 5　自回归生成 74
实验 6　模型评估 76
实验 7　模型安全 78
实验 8　上下文学习 80
实验 9　检索增强及智能体 82
实验 10　预训练工程 84
实验 11　后训练方法 86
实验 12　参数高效微调 88
实验 13　训练数据生产 92
实验 14　推理训练 94
实验 15　资源规划 96
实验 16　训练系统优化 98
实验 17　推理服务 101
实验 18　推理优化 105

实验 19 模型压缩 . . . 107
实验 20 掩码语言建模 . . . 109
实验 21 开放模型制品审计 . . . 111
实验 22 视觉 Transformer . . . 113
实验 23 多模态数据 . . . 115
实验 24 条件视觉生成 . . . 117

第三部分 参考解析 119

第一章 绪论 . . . 120
第二章 数学基础 . . . 120
第三章 神经网络基础 . . . 121
第四章 文本表示 . . . 123
第五章 注意力机制 . . . 125
第六章 Transformer 网络结构 . . . 126
第七章 位置表示 . . . 128
第八章 自回归语言建模 . . . 129
第九章 模型评估 . . . 130
第十章 模型安全 . . . 132
第十一章 上下文学习 . . . 134
第十二章 信息检索 . . . 136
第十三章 检索增强生成 . . . 138
第十四章 领域应用 . . . 140
第十五章 智能体架构 . . . 141
第十六章 智能体运行系统 . . . 143
第十七章 优化及泛化 . . . 145
第十八章 训练数据工程基础 . . . 146
第十九章 训练稳定性 . . . 148
第二十章 监督微调 . . . 150
第二十一章 参数高效微调 . . . 151
第二十二章 训练工程优化 . . . 152
第二十三章 预训练原理 . . . 153
第二十四章 强化学习基础 . . . 154
第二十五章 人类反馈学习 . . . 156
第二十六章 可验证推理训练 . . . 157
第二十七章 模型规模及资源预算 . . . 159
第二十八章 算力基础设施 . . . 160
第二十九章 分布式训练 . . . 162

第三十章　增量解码 163
第三十一章 批处理调度 164
第三十二章 模型服务架构 166
第三十三章 推理计算优化 168
第三十四章 模型优化 169
第三十五章 双向语言建模 173
第三十六章 混合专家模型 174
第三十七章 长上下文建模 175
第三十八章 开源模型体系 176
第三十九章 视觉 Transformer 178
第四十章　多模态建模 179
第四十一章 扩散模型的概率原理 181
第四十二章 条件视觉生成 183

第四部分　参考文献 **185**

第一部分

习题

第一章　绪论

1.1. 在本章两步生成例子中，若将 $p(u \mid b)$ 改为 0.7，最大联合概率路径是否仍经过 b？写出全部路径概率后判断.

1.2. 同一段文本被两个分词器编码为不同长度，为什么不能仅比较每词元生成速度就断定哪个系统处理文本更快？

1.3. 一个知识问答系统检索到了正确文档，却给出了错误答案. 应检查哪些环节？为什么仅扩大模型不能保证解决问题？

1.4. 某次修改使训练吞吐提高，但有效监督词元减少. 需要补充哪些证据才能判断这是否是同一学习目标下的加速？

第二章　数学基础

2.1. 保持验证器的接受率设定不变，将先验正确率从 0.2 改为 0.8，计算通过后的正确率并解释变化.

2.2. 设 $p = (1, 0)$，分别计算其对 $q = (\frac{1}{2}, \frac{1}{2})$ 的交叉熵，以及反向 KL 散度. 哪些量有限?

2.3. 对 $p = (0.6, 0.4)$、$q = (0.7, 0.3)$，验证交叉熵分解并比较两个方向的散度.

2.4. 将一个含四个目标的文档切成两个等长片段. 什么条件下总词元损失保持不变? 片段切断上下文后为什么未必保持?

2.5. 若一个模型将全部分数同时加上 1000，其理论概率应怎样变化? 朴素实现可能发生什么数值问题?

2.6. 设给定输入的真实正类概率为 $p \in (0, 1)$，模型输出为 $q \in (0, 1)$，正类权重 $w > 0$. 对条件期望损失 $L(q) = -wp \log q - (1 - p) \log(1 - q)$，求最优输出 q^*，并推导从 q^* 恢复 p 的反变换. 取 $p = 0.1, w = 9$ 计算 q^*，解释为何提高少数类权重后不能直接把输出当作真实概率，并说明反变换的适用前提.

第三章　神经网络基础

3.1. 设 $X \in \mathbb{R}^{3\times 4}$、$W \in \mathbb{R}^{4\times 2}$、$b \in \mathbb{R}^{2}$. 写出 $Z = XW + \mathbf{1}_3 b^{\mathsf{T}}$ 的分量式，并给出上游梯度为 G 时三个输入梯度的形状. 解释为什么偏置梯度不是 G 本身.

3.2. 从矩阵内积定义出发，推导 $\mathcal{L} = \dfrac{1}{2}\|XW - T\|_F^2$ 对 W 和 X 的梯度，其中 T 为固定目标矩阵. 若损失改为对样本数取平均，哪些项需要改变?

3.3. 对 $q = (\dfrac{1}{2}, \dfrac{1}{3}, \dfrac{1}{6})$ 写出完整 Softmax 雅可比，验证各行之和为零，并用一个给定向量检查显式矩阵乘积与向量形式相等.

3.4. 将目标改为权重向量 $y = (1, 1, 0)$，推导 $-\sum_j y_j \log q_j$ 的分数梯度. 说明它为何不再等于 $q - y$，以及它与三个独立二元分类目标有何区别.

3.5. 设 $X \in \mathbb{R}^{2\times 3\times 4}$，偏置形状为 $1 \times 3 \times 1$. 写出 $Z = X + b$ 的反向求和轴. 若前向又对全部 24 个元素取平均，说明归一化系数在反向中的位置.

3.6. 嵌入编号为 $(1, 3, 1, 3)$，四个位置的上游梯度依次为 $(1, 0)$、$(0, 1)$、$(2, 3)$、$(-1, 2)$. 求嵌入表各行的梯度，并说明按最后一次出现覆盖会造成什么错误.

3.7. 在本章两层网络例题中，单独扰动 $W_{1,12}$. 保持 ReLU 分支不变，写出标量损失关于扰动的表达式，并核对其零点导数为 $\dfrac{a}{2} + b$.

3.8. 对 $f(x) = \max(0, x)$ 在 $x = 0$ 使用中心差分. 所得值是多少? 它为什么不能用于否定零点取零的求导约定? 再解释在 $x = 1$ 附近如何选择扰动才能避开折点.

3.9. 一个长度不同的序列批次用填充后的总位置数作交叉熵分母，另一个实现用有效词元数作分母. 分别写出两种目标，分析它们在填充比例变化时的梯度尺度. 即使二者均通过有限差分校验，是否足以认定目标相同?

3.10. 某语言模型的自动微分梯度与中心差分一致，但训练目标把每个位置的当前词元当作标签. 说明这项检查能证明什么；构造一个长度为 3 的词元序

列，写出正确与错误的输入标签配对，并设计独立于梯度差分的检查．讨论在 ReLU 折点附近做差分时应如何解释不一致．

第四章　文本表示

4.1. 对训练片段 `aaaa`（权重为二）与 `aab`（权重为一），计算第一轮全部相邻对计数，按从左向右的非重叠替换规则执行最高频合并．比较频数与实际词元减少量，并解释差异．

4.2. 使用本章三轮 BPE 规则编码 `abacabab`．写出各轮边界；再人为交换前两条规则，判断此输入是否一定改变．说明“某个例子相同”为什么不足以证明规则顺序无关．

4.3. 构造一个 WordPiece 词表，使最长前缀选择后无法完整覆盖输入，但回溯搜索可以找到完整切分．写出两条路径，并说明为何不能将回溯结果归因于原始贪心编码器．

4.4. 用 Unigram 例题的第一组参数，枚举 `abab` 的全部分段，分别计算总权重和最优路径权重，再用前向动态规划复核．说明两者相等需要什么条件．

4.5. 对字符串 `aba`，根据本章 EM 更新后的参数重新计算三条后验概率．判断仅用这一条训练样例是否足以形成适用于其他语言的词表，并解释原因．

4.6. 证明：若规范化函数不是单射，且不附带额外恢复信息，则不存在能对所有原文无损恢复的解码器．给出代码或数学文本中不宜随意合并的两种字符形式．

4.7. 词表翻倍且序列平均长度下降 20% 时，分别计算密集注意力与全词表输出投影的相对规模．列出仍然影响总时延的至少三个因素．

4.8. 为“提示两个词元、回答两个词元、真实 EOS、填充两个位置”的样本写出显式右移后的输入、目标、输入有效性与仅回答损失掩码．令 PAD 与 EOS 共享编号，指出哪些位置仍应计算损失．

4.9. 两套分词器对测试文本均能无损解码，但输出编号不同．解释直接替换分词器可能破坏哪些模型参数与服务资产，并设计能够定位差异层次的比较顺序．

4.10. 对本章的三词元查表示例，直接写出独热向量与矩阵乘法结果．若词表第 1 与第 3 行交换而输入编号保持不变，哪些输出发生变化？如果同时调整编号映射，结果又如何？

4.11. 在式 (4.47) 中令 $\alpha = 2$．说明为什么只有输入路径梯度显式带有该系数，而输出路径不能机械乘二．讨论 H 的数值仍可能怎样随输入缩放变化．

4.12. 取两个向量 $(1, 1)$、$(1, -1)$，再将第一维都放大十倍．比较变换前后的内积和余弦相似度，解释一般可逆坐标变化为什么不能保持嵌入几何．

第五章　注意力机制

5.1. 设 $Q \in \mathbb{R}^{5\times 8}$、$K \in \mathbb{R}^{7\times 8}$、$V \in \mathbb{R}^{7\times 3}$，写出分数、权重与输出的形状. 为什么不能沿查询轴做本章定义的归一化?

5.2. 当查询和键各坐标方差均为 4 且满足独立零均值条件时，推导未经缩放与除以 $\sqrt{d_k}$ 后的分数方差. 为什么缩放后不一定是单位方差?

5.3. 对本章单头例题，去掉因果掩码，计算第一行权重及输出. 再将第三个值向量替换为零，比较第一行输出变化.

5.4. 证明 Softmax 雅可比满足 $J\mathbf{1} = 0$，并解释这一性质与分数常数平移不变性的关系.

5.5. 对第二查询的局部损失 $\mathcal{L} = o_{2,1}$，手算 $\overline{K}_1, \overline{K}_2$. 用中心差分验证其中一个坐标，明确扰动时哪些变量保持固定.

5.6. 若唯一查询绝对位置是 8，键位置为 0 至 8，写出正确因果允许集合. 说明直接使用 1×9 左上对齐下三角掩码会遗漏哪些位置.

5.7. 设 $H_q = 24, H_{kv} = 6, d_h = 64$，一个样本缓存 2048 个位置，每元素两字节，计算一层键值存储量，以及与 MHA 的比例.

5.8. 若两个头的 A 相同而 V 不同，它们的输出是否必然相同? 若两个头的 K, V 相同而 Q 不同呢? 分别给出反例或证明.

5.9. 证明权重 Dropout 后行和的期望仍为一，并计算独立掩码下该行和的方差. 由此解释为什么训练时观测到行和偏离一不必然表示 Softmax 实现错误.

5.10. 构造一个全屏蔽的合法查询与一个仅用于填充的无效查询，分别说明应如何处理. 为什么把分数设为统一的大负数不是正确通用方案?

5.11. 固定查询头数 $h_q = 32$、头维度及序列长度，将键值头数由 32 改为 8. 分别判断键值投影输出、KV 缓存和注意力分数数量如何变化. 解释为何缓存缩小四倍不能推出端到端时延缩短四倍，并列出验证实际收益应固定的条件.

第六章　Transformer 网络结构

6.1. 证明不含激活的两层 FFN 可合并为一个仿射变换．加入 ReLU 后，写出给定激活区间内的局部雅可比，并解释为何跨区间不再是同一个仿射函数．

6.2. 对 $u = -1, 0, 1$ 计算 GELU、SiLU 及其导数，比较二者在负区间的行为．说明“平滑”为什么不足以推出训练效果更好．

6.3. 设普通 FFN 宽度为 $f = 4096$，求等无偏置参数预算的门控宽度．若宽度必须是 256 的倍数，分别给出向下、向上取整后的参数差异．

6.4. 从 $c = Px$ 出发独立推导 LayerNorm 雅可比．验证该雅可比作用于全一向量为零，并说明这与平移不变性的关系．

6.5. 对 $x = (1, 2, 3)$ 分别计算无仿射 LayerNorm 和 RMSNorm．给向量整体加十，比较输出变化；不要把非零 ϵ 情况下的比例缩放近似写成精确不变性．

6.6. 取一维 $F(x) = ax$，分析十层残差网络在 $a = 0.1, -0.1, -1$ 时的导数．结合归一化雅可比，解释恒等路径不能独立保证稳定训练．

6.7. 写出长度六、条件前缀长度三的前缀可见性矩阵．若对前缀内部使用普通下一词元监督，指出可能出现泄漏的具体位置．

6.8. 对 $d = 768, f = 3072$ 计算带偏置编码器层与含交叉注意力解码器层的参数量，列出未包含的模型组件．再将普通 FFN 改为等预算无偏置 SwiGLU，重新计算．

6.9. 将完整例题的因果掩码改为双向掩码，逐步重算输出．解释为什么此时位置一可能依赖位置二；再讨论二维末端 LayerNorm 可能隐藏哪些数值差异．

6.10. 目标嵌入与输出头共享权重时，说明一个未出现在输入中的词元行为什么仍可能获得梯度．设计重载检验，同时覆盖共享关系、固定位置状态和生成首个词元．

6.11. 将手写 Transformer 迁移到库实现后，参数均能加载且张量形状一致，但首步 logits 不同．设计从输入到输出的逐层对照方案，至少覆盖掩码语义、归一化位置、位置表示和权重布局，并说明何时应开始比较完整生成结果．

第七章　位置表示

7.1. 写出交换长度为 3 的序列前两行的置换矩阵 Π，验证 $\Pi^{\mathsf{T}}\Pi = I$．设 M 为下三角可见掩码，比较 M 与 $\Pi M \Pi^{\mathsf{T}}$ 的允许位置.

7.2. 证明逐位置函数 f 满足 $f(\Pi X) = \Pi f(X)$；再证明对输出按位置求和会将这种等变性变为排列不变性．解释两者为何不能使用同一个术语.

7.3. 对 $d = 4$ 的正弦位置向量，计算 $p_0^{\mathsf{T}} p_2$，再用频率差公式独立核对．验证 $\|p_t\|_2^2 = 2$.

7.4. 正弦编码采用 $(\sin\theta, \cos\theta)$ 的分量顺序，RoPE 采用标准列向量旋转矩阵．比较式（7.8）与式（7.13）中的正负号，说明为什么不能不看坐标约定就直接互换.

7.5. 用 $q = (1, 2)^{\mathsf{T}}$、$k = (3, -1)^{\mathsf{T}}$ 和频率 1，计算位置 $(i, j) = (0, 2)$、$(1, 3)$、$(6, 8)$ 的旋转内积．再计算 $(3, 1)$，分析位置差反号的作用.

7.6. 若仅旋转四维向量的前两维，推导完整内积公式，并给出一个数值例子，说明未旋转部分为何不受位置差影响.

7.7. 某实现每一步都将新查询的位置设为零，但历史键保留原位置的旋转．写出它实际使用的相对角度，并与正确角度比较．说明为什么程序能够输出合法张量却依然违反模型定义.

7.8. 有人声称“正弦函数可计算任意位置，所以不需要长文本训练”“RoPE 的距离越远，点积必然越小”．分别给出反驳所需的数学边界或反例，并说明还需要什么任务证据才能讨论泛化.

第八章　自回归语言建模

8.1. 对序列 $(\mathrm{BOS}, a, b, \mathrm{EOS})$，分别写出数据侧移位与损失侧移位的输入、目标和有效位置．说明两个接口连续使用会把目标变成什么．

8.2. 给定 $p(a \mid c) = 0.7$、$p(b \mid c, a) = 0.5$、$p(\mathrm{EOS} \mid c, a, b) = 0.8$，计算完整响应 (a, b) 的概率、负对数似然及按三个目标平均的损失．若遗漏 EOS，比较对象发生了什么变化？

8.3. 对式（8.10）证明正温度保持排序．最大分数并列时，为什么温度趋零的极限与某个确定并列规则的贪心输出不完全相同？

8.4. 对概率 $(0.4, 0.3, 0.2, 0.1)$，计算阈值 $\rho = 0.5$ 与 $\rho = 0.95$ 的核集合及最终分布．指出“累计超过阈值的项全部移除”会造成的错误．

8.5. 用图（8.3）计算四条路径的对数分数，并分别模拟宽度一与宽度二的算法（8.2）．明确并列规则与已完成候选是否继续占据束位置．

8.6. 说明原始对数概率为何具有扩展上界．再构造一条前缀，使长度归一化后的分数在继续生成后反而上升．

8.7. 若每步 EOS 概率为固定 $\epsilon \in (0, 1)$，推导生成步数的分布与期望．随后解释将 EOS 永久排除出 Top-k 集合会怎样改变终止性质．

8.8. 某批次有三个请求，其中一个已命中 EOS．说明其后应如何管理完成状态、填充位置与缓存，为什么不能仅让它继续生成后再删去文本．

8.9. 构造一个跨两个词元解码边界的停止串，说明逐词元独立字符串比较为何可能遗漏它．

8.10. 教师强制精确对应序列最大似然，为什么模型仍可能在自由生成中退化？分别从有限数据、解码变换和终止规则举例，避免将它们统称为暴露偏差．

8.11. 给定证据只说明“系统于 2024 年发布”，回答却称“系统由甲于 2023 年发布”，并附上一个证据包中不存在的条目编号．分别判定其中的事实错误、上下文矛盾、无依据生成和虚构归因；再说明降低温度为何不能修复这些问题，并给出至少两项系统级控制．

第九章　模型评估

9.1. 由二分类混淆计数推导 F1 的计数形式，并构造准确率高但正类召回率很低的例子.

9.2. 证明完整类别集合下单标签分类的 Micro-F1 等于准确率，并给出该等式在多标签任务中失效的例子.

9.3. 若正确决策也有成本，推导替代式（9.6）的阈值，并说明何时应增加弃权行动.

9.4. 构造四个按预测分数排序的标签，使精确率在提高阈值时先升后降；解释召回率为何没有相同现象.

9.5. 用不同分词方式表示同一段文字，说明跨分词器 PPL 的计量问题. 若只计回答词元，应如何定义掩码和分母?

9.6. 为本章词面算例添加一个重复词，重算 Token F1 与一元截断精确率，并分析未经截断会产生什么偏好.

9.7. 从子集计数推导式（9.9），说明相关采样与完美候选选择假设分别影响哪一层结论.

9.8. 设计裁判顺序、长度和家族偏好的正交对照，说明总体人工一致率为什么不能识别全部偏差.

9.9. 由二次不等式重新推导 Wilson 区间，计算全失败情况下的上界，并与零失败的单侧精确界比较.

9.10. 对用户平均和请求平均各设计一组大小不等的簇算例，使两种总体指标明显不同.

9.11. 对“同一 100 题、五个种子”的结果，分别说明固定案例集运行稳定性与新案例泛化需要哪些随机性假设.

9.12. 某实验每天检查一次显著性，在首次通过时停止. 解释为什么固定样本检验的标称水平不能直接沿用，并提出预先可执行的替代协议.

9.13. 为一个检索问答系统定义质量、错误损失、延迟和成本的可行域，构造三个互不支配配置，并说明业务权重应在哪一步介入.

9.14. 一个候选总体得分提高，但高风险切片样本数很少. 写出可支持与不可支持的结论，以及下一轮最有信息量的抽样方案.

9.15. 比较表（9.2）的两行 OSWorld 部分得分. 列出无法据此直接宣布哪一个模型更强的具体协议差异；再设计可比较的受控实验.

9.16. 一条提示含五项约束，满足其中四项. 分别说明 IFEval 提示级与指令级计分的差别，并解释为什么格式通过不能代替事实正确.

9.17. 为中文企业知识问答选择一组基础、检索及端到端评测. 写明每项基准无法覆盖的业务要求，以及新增保留集的来源.

第十章　模型安全

10.1. 为一个仅回答问题的模型和一个可修改业务记录的模型分别列出资产、不可接受后果和新增信任边界.

10.2. 推导期望损失公式在互斥场景与可加损失分量两种定义下的含义，构造攻击路径重叠导致重复计数的例子.

10.3. 使用不同单调编码展示序数风险乘积排序可改变，说明为什么它不能解释为真实风险倍数.

10.4. 两层控制的漏放概率分别为 0.2 和 0.3，给出联合漏放率的可行范围，并构造独立与相关失效情形.

10.5. 对内部报告外发场景，分别指出模型行为控制、授权控制、凭据范围与输出渲染能够阻断的路径.

10.6. 设计一个规范化资源的例子，说明先检查字符串、后解析资源可能怎样破坏授权对应关系；只讨论防御契约，不执行攻击.

10.7. 为需要批准的动作设计摘要字段，说明参数、权限或资源版本改变后如何判定原批准是否仍有效.

10.8. 工具已经执行但响应超时，设计执行账本和对账状态机，说明为何使用新幂等键重试可能错误.

10.9. 比较哈希一致、签名可信、加载成功和行为评测通过分别能够支持什么结论.

10.10. 设计样本到数据快照、模型和部署的谱系，说明删除原始数据后仍需分析哪些派生资产.

10.11. 根据不同危险请求基率计算拦截精确率，解释为什么同一过滤器在不同业务中审核负担不同.

10.12. 从零失败概率推导单侧置信上界；若要在独立同分布前提下将 95% 上界降至 1% 以下，推导最小样本数.

10.13. 为多语言安全评测设计正常、风险和无正常样本三类切片，明确定义漏放、误拦及控制覆盖的分母.

10.14. 将构造事件扩展为结果未知而非已成功的工具调用，说明调查、遏制与恢复流程中哪些判断必须改变.

第十一章　上下文学习

11.1. 用条件概率说明：改变少样本示例与执行一次参数微调有何不同？为什么零样本不代表模型从未学习过该任务？

11.2. 为式（11.2）设计一个包含相关性、技能覆盖和冗余的具体评分，说明评分可以使用哪些信息、不得使用哪些测试信息.

11.3. 给出两个示例排列，使最后出现的标签不同．设计能够区分顺序效应与示例内容效应的成对实验.

11.4. 对三个标签概率 $(0.5, 0.3, 0.2)$ 与偏好估计 $(0.6, 0.2, 0.2)$ 计算上下文校准结果，并解释为何均衡标签未必符合真实任务先验.

11.5. 构造五条路径，使最大概率路径的答案不同于边缘概率最大的答案，完整列出两种决策.

11.6. 证明独立采样频率估计的均值和方差，并说明温度采样改变了哪个分布.

11.7. 对 $p = 0.4, K = 5$ 计算二元多数正确概率．再构造三个答案的分布，使正确答案概率仍为 0.4 却是唯一众数.

11.8. 五条候选产生 $2:2:1$ 时，为什么奇数候选不能排除平票？设计不依赖标准答案的处理规则.

11.9. 区分候选覆盖率 $1-(1-p)^K$、最高票正确率和验证选择正确率，并说明三者的前提.

11.10. 给定当前票数 $5:2:1$、最多还有两条候选，判断是否可以保持最高票决策不变地提前停止．若还有四条会怎样？

11.11. 修改综合算例，使一条错误候选通过算术执行但违反题意，说明验证器还应检查哪些条件.

11.12. 在固定预算下比较增加示例与增加候选的取舍，分别考虑无前缀缓存、共享前缀以及并行生成三种执行条件.

11.13. 为结构化答案定义单位、等价表达、缺失值和解析失败规则，并分别写出总体支持率与条件支持率.

11.14. 设计提示迁移报告，说明更换模型后哪些结论必须重新评价，以及如何避免用多个候选夸大测试样本数.

第十二章　信息检索

12.1. 为带表格、版本日期和访问限制的文档设计片段记录，说明哪些上下文必须随片段保留.

12.2. 从二元条件独立假设推导词项对数证据，并指出它与完整 BM25 公式之间还需要哪些建模选择.

12.3. 对式（12.3）计算词频因子的两个导数，分析 $b = 0$ 与 $b = 1$ 的含义.

12.4. 改变评分例题中文档 B 的长度，求使其得分高于 A 的一个长度，并解释其他条件是否保持一致.

12.5. 推导对比损失对查询向量的梯度；如果某个高分负例其实相关，梯度会造成什么后果?

12.6. 构造两个未归一化文档向量，使内积、余弦和欧氏距离产生至少两种不同排名.

12.7. 解释为什么 HNSW 增加查询探索宽度可能提高召回，却不能对所有过滤条件保证精确结果.

12.8. 给定 $h = 768, m = 48, k = 256$，计算每条 PQ 码的字节数及全部子码本中心的元素数，不含原始精排向量.

12.9. 推导 PQ 距离误差界，并说明候选间距很小时为何容易改变排名.

12.10. 使用本章两路排名，分别取 $c = 1$ 和 $c = 60$ 计算 RRF，比较顶部差异的影响.

12.11. 说明为什么重排器不能补救候选漏召回，以及增加候选深度会消耗哪些资源.

12.12. 设计文档撤销后在倒排索引、图索引、向量缓存和检索缓存中的删除传播流程，区分立即不可见与物理清理.

12.13. 为前四名等级 $(1, 0, 3, 2)$ 计算 DCG 与 nDCG，并与完整算例比较 Recall、MRR 的不同关注点.

12.14. 设计一个 ANN 召回很高但任务召回很低的例子，说明应改索引还是改表示.

第十三章　检索增强生成

13.1. 从序列潜变量和逐词元潜变量的定义分别推导两种 RAG 分解，说明为什么交换乘积与求和一般不成立.

13.2. 给出三个文档的检索概率及二词元条件概率，自行计算两种边缘概率；再截断至前两篇，比较保留原概率质量与局部重归一化.

13.3. 推导候选集合固定时检索逻辑值的梯度 $p_i - \gamma_i$. 若候选遗漏唯一有效证据，解释这个梯度不能解决什么问题.

13.4. 对重排器的成对逻辑损失求正负候选分数的偏导，并说明为什么候选采样分布影响训练目标.

13.5. 在本章 320 词元例题中，为 d_2 增加 40 词元的必要定义，重新求满足预算且覆盖尽可能多事实的组合. 写明并列最优时的选择规则.

13.6. 为含有历史日期和对话指代的问题设计两条改写，列出不得变化的约束，并构造一个检索分数提高却问题语义改变的反例.

13.7. 将“某负责人所管理服务的故障数”分解为有依赖关系的查询图，说明负责人变更与数据快照怎样进入数据契约.

13.8. 构造一句带有真实引用 ID 但来源不支持的答案，分别说明格式、定位和语义检查的结果.

13.9. 在拒答成本模型中令答对也有成本 C_c，推导新的回答条件，说明该模型在哪些情况下不足以表达部分回答.

13.10. 设计文档更新、权限撤销和新文档加入三种缓存失效路径，证明只检查已引用文档的版本为何不足以处理负缓存.

13.11. 针对重复和乱序摄取事件设计幂等键与版本比较规则；源系统没有全局版本号时，说明能保证和不能保证的一致性.

13.12. 给出三层条件成功率，并计算完整支持路径的成功率. 再构造一个缺证据但凭参数记忆答对的情形，解释两个指标的差别.

13.13. 设计包含可回答、不可回答、权限受限与上游异常的评估集，分别定义正确回答率、过度拒答率和故障率的分母.

13.14. 选择一种重排优化，设计包含充分证据替换、证据删除与版本冲突的反事实评估，说明如何控制生成随机性及预算.

第十四章　领域应用

14.1. 将“提升客服效率”改写为可计数任务，定义输入、输出、结束条件与失败.

14.2. 为合同条款抽取设计证据记录，说明为何仅保存文件名不足以复核.

14.3. 对库存查询列出身份、单位、状态与时间字段，并构造遗漏任一字段导致错误的例子.

14.4. 根据式 (14.2) 计算辅助操作 5 分钟、复核 4 分钟、30% 任务返工 8 分钟时，相对 12 分钟原流程的净收益.

14.5. 将自动与人工两种决策都加入执行成本及残余错误，重新推导自动化阈值.

14.6. 为制造业告警分析区分观察、统计关联、故障假设与已确认原因.

14.7. 为什么教育对话满意度不能直接证明学习效果？设计更贴近目标的测量.

14.8. 给出一项代码任务，说明编译、测试、评审与真实部署分别证明什么.

14.9. 构造上线前后成功率改善但实际处理能力未改善的任务分布变化.

14.10. 设计以设备为分组单位的评估划分，解释随机划分日志行的泄漏风险.

14.11. 在自动覆盖算例中，将所有自动错误交给人工返工，重新定义干预率与成本分母.

14.12. 为一个领域应用制定可复现失败记录，明确哪些敏感信息不需要保存原文.

14.13. 制造业知识助手根据维修文档给出高置信故障判断，并建议调整设备控制参数. 设计从证据检索到允许执行动作的流程，区分模型置信度、操作权限和独立安全约束，并说明审计记录应包含哪些信息.

第十五章　智能体架构

15.1. 从贝叶斯公式推导式 (15.1)，解释分母和动作对状态转移的影响. 自然语言摘要在什么条件下会丢失决策必需的信息?

15.2. 对同一文档查询任务分别设计固定工作流和受限 ReAct；指出需要动态决策的观测，而不只增加模型调用次数.

15.3. 为具有三个依赖子任务的计划写出前置条件、完成判据和版本变化后的重规划规则.

15.4. 证明只有步骤上限不能限制墙钟时间，只有工具次数上限不能限制无工具决策循环，并给出同时满足两类约束的终止设计.

15.5. 将数例中的查询成本改为 0.8，计算信息价值；再说明查询后状态变化为何需要引入额外转移假设.

15.6. 为一条配置记忆设计来源、有效时间、版本与冲突字段；解释观察时间晚于另一条记录为何不必然意味着它更权威.

15.7. 推导恒定失效风险率下的新鲜度衰减，并举出两个不适合使用该模型的记忆类型.

15.8. 在共同历史重读模型中，令 $n = 4, m = 100, R = 3$，计算输入词元总量；设计差量消息机制并说明其遗漏上下文风险.

15.9. 三个协作者拥有相同错误来源时，多数表决为何不提供独立判断的收益? 设计一个区分角色多样性与证据多样性的评估.

15.10. 两个协作者同时更新版本 12 的任务状态，写出采用 CAS 的成功及冲突路径. 哪些数据可以追加合并，哪些必须原子校验?

15.11. 为“远端预留成功、本地回执丢失”写出恢复状态转移，区分确认未提交、结果未知、幂等重试和补偿.

15.12. 设计预算固定的单智能体与监督者架构比较，给出任务完成、副作用、通信、迟到结果和终止原因的评价口径.

15.13. 同一模型在两个 Harness 中成功率不同. 前者有两倍工具调用预算且允许访问额外数据，能否将差异解释为 Harness 的控制逻辑更优？请设计可归因的比较.

第十六章　智能体运行系统

16.1. 为知识回答、操作提案和已完成写入分别设计输出契约，并说明各自需要什么证据.

16.2. 证明确定转换函数按同一事件序列重放得到相同状态；加入一次未记录的外部读取后，该证明在哪一步失效?

16.3. 给出两个工作进程同时更新任务的时序，说明版本条件如何避免丢失更新，以及为何仍需外部隔离机制.

16.4. 构造服务端先登记键再执行与先执行再登记的两种崩溃窗口．怎样的原子边界可以消除重复副作用?

16.5. 发件箱消息已发送但发送状态未更新时进程崩溃，恢复后会发生什么？为什么不能因此删除接收方去重?

16.6. 推导预算不变量，并讨论工具最终成本超过预留时应如何改变契约.

16.7. 为“创建工单成功但响应丢失”设计状态机，包括自动核对与人工核对路径.

16.8. 设计一个审批摘要，证明参数变化不能复用旧审批；讨论对象版本变化但参数不变的情况.

16.9. 对不可逆通知设计补偿策略，解释哪些事实无法撤销.

16.10. 区分轨迹回放、真实接口集成验证和线上任务评估的证据范围.

16.11. 构造两个强相关失败环节，使边际成功率乘积显著偏离端到端成功率.

16.12. 设计旧任务跨应用版本升级的恢复方案，列出必须保留的解释逻辑与版本字段.

16.13. 一张 1920 乘 1080 的完整视口截图对应 960 乘 540 的逻辑坐标视口，原点为零；目标位于截图坐标 (1200,600)．计算动作坐标，并说明页面滚动后能否直接复用.

16.14. 某界面任务有 10 步，每步在此前均成功的条件下成功概率都是 0.98．求全程成功率；提交最后一步后回执丢失，是否应从头重放？

16.15. 比较 CDP、Computer Use 和智能体工具协议所在的层次．通过一个工具协议连接浏览器后，是否自动获得跨浏览器兼容和安全授权？

第十七章　优化及泛化

17.1. 对 $F(u,v)=\dfrac{2u^2+8v^2}{2}$，求梯度下降对任意初值收敛的学习率区间，以及平衡最坏方向收缩因子的学习率．解释边界取值为何不能保证收敛．

17.2. 对式（17.11），证明目标值满足 $F(\theta_k)\leqslant \rho^{2k}F(\theta_0)$．指出正定假设在证明中的作用．

17.3. 某总体的单样本梯度协方差为 Σ．在有放回独立抽样下将批量从 B 增加到 $4B$，梯度标准差如何变化？若每个批量内样本梯度完全相同，为什么这一结论不再成立？

17.4. 从式（17.23）求一维二次目标的稳态期望损失．解释固定批量时降低学习率的收益及其代价．

17.5. 对负半轴斜率为 0.1 的分段线性激活，推导输入宽度为 256 时的前向初始化方差．说明“保持二阶矩”与“每次具体初始化均保持范数”的区别．

17.6. 在式（17.39）中改用平方损失的一半，写出对应的惩罚系数．若掩码不独立，需要加入什么交叉项？

17.7. 若对目标函数整体乘以正常数 c，固定学习率的梯度下降会发生什么变化？如何同时调整学习率和正则系数，保留原来的数据项与惩罚项比例及更新轨迹？

17.8. 某公司按记录随机划分同一客户的多次交易，并在全部记录上拟合标准化器．分别指出两类泄漏，并为“预测未见客户的未来交易”设计训练、验证和测试边界．

17.9. 推导收缩例题中使偏差平方与预测方差之和最小的 a，说明为什么计算出的最优值不能直接作为未知总体上的可用超参数．

17.10. 一个模型的训练损失与验证损失都很高，开发者准备增强正则化．设计检查顺序，区分损失统计口径不一致、数据或标签错误、优化失败和容量不足；说明哪些证据才支持过拟合判断，以及小批次拟合检查的局限．

第十八章　训练数据工程基础

18.1. 某数据集每组有十条观测，组内相关系数为 0.2. 在等组大小假设下推导均值方差与等效样本量，并说明该量为何不能直接视为神经网络训练的有效样本数.

18.2. 为含工具调用、图片引用和偏好对的数据设计公共信封与三种载荷，列出必须由模式检查的跨字段约束.

18.3. 推导式 (18.3)，构造两个领域原始数量相同但过滤后严重失衡的例子，并说明校正所需的信息.

18.4. 对 $J = 0.5$，希望 MinHash 估计量标准差不超过 0.025，在独立排列模型下至少需要多少行？该条件是否保证每次误差都小于 0.025?

18.5. 固定签名长度 $m = 100$，比较 $(b, r) = (20, 5)$ 与 $(10, 10)$ 对相似度 0.6 的候选概率，解释计算量与漏检的取舍.

18.6. 用本章三个集合说明：为什么近重连通分量适合保守划分，却不能无条件作为“全部内容可互相替代”的证明?

18.7. 新数据把已分到训练和测试的两个家族连接起来. 提出一个保留既有评估版本解释性的处理方案，明确哪些记录被隔离、哪些版本必须失效.

18.8. 三个来源的记录数为 $(10000, 1000, 100)$，平均有效长度为 $(50, 200, 500)$. 计算 $\alpha = \dfrac{1}{2}$ 下记录概率与近似词元份额，并给出面向指定词元目标的调整思路.

18.9. 对长度 $(6, 6, 6)$ 与块长 10，比较连续流、不可拆原子样本以及独立注意力打包的容量和语义，不能仅比较填充比例；独立注意力打包分别讨论允许与禁止跨块拆分的条件.

18.10. 设计一个在块中跨越两个来源的词元跨度表，包含插入分隔符、填充与标签屏蔽. 说明如何从一个输出位置反查到原始记录.

18.11. 某作业先提交训练检查点，后保存数据游标，在两步之间崩溃．分析恢复后可能出现的重复，并给出一致性提交所需的状态边界．

18.12. 一个源文档已从对象存储删除，但被去重代表、派生分片和训练检查点引用．画出需要处理的血缘图，并区分未来数据撤销与既有模型影响．

第十九章　训练稳定性

19.1. 展开平方梯度的指数平均，并在各步二阶原点矩相同的假设下推导偏差修正.

19.2. 在两步算例的第二步之前把一阶矩、二阶矩和优化器内部步号一起清零. 为避免零梯度坐标的 $\frac{0}{0}$，采用正 ϵ 并报告 $\epsilon \to 0^{+}$ 的极限轨迹，解释与原轨迹的差异.

19.3. 证明范数裁剪是欧氏球投影，并给出逐坐标裁剪改变方向的例子.

19.4. 推导非零 ϵ 下缩放整个梯度序列对 Adam 更新的影响.

19.5. 设两组参数梯度范数为 3 和 4，求阈值 2 时的全局裁剪结果，比较分别裁剪两组的结果.

19.6. 为不使用预热的情况定义无除零、端点明确的余弦日程，并写出首末更新的学习率.

19.7. 说明梯度累积中途保存检查点相比更新边界保存多需要哪些状态.

19.8. 数据加载器已预取但尚未消费三个批次. 设计恢复方案，避免重复计入消费量或遗漏批次.

19.9. 为分片缺失、摘要错误和资产版本不匹配分别设计恢复拒绝条件.

19.10. 解释为何完整检查点不保证改变设备数后的逐位复现，并区分语义和数值验收.

19.11. 训练出现非有限梯度，某次更新可能已经改变优化器状态. 设计覆盖更新提交前后两种情况的处理流程，说明降低学习率、梯度裁剪与损失缩放各自能解决什么，并列出恢复时必须保持一致的状态和计数.

第二十章　监督微调

20.1. 从联合概率分解推导全序列损失与回答条件损失的差异. 给出一个有理由监督提示的任务，并说明代价.

20.2. 在本章移位算例中，将回答扩为三个词元并在尾部共设置两位填充，列出 r_j、m_t 与全部有效预测配对.

20.3. 若 PAD 与 EOT 使用同一词元编号，设计一个不会误删真实终止监督的位置判定规则.

20.4. 两个样本的回答长度为 4 与 12、平均损失为 2 与 1，分别计算词元等权、样本等权与序列损失，并解释梯度权重.

20.5. 对三轮对话展开全部前缀，并在每个前缀监督所有助手轮次，求各轮被计权的次数；给出每轮恰好一次的构造方案.

20.6. 利用式（20.9）解释提示位置标签被忽略时，其嵌入参数为什么仍可能更新.

20.7. 候选正确率为 0.7，过滤器正确接受率为 0.8，错误接受率为 0.1，计算保留率和保留集合正确率，并讨论误差相关性的影响.

20.8. 设计多轮数据的一致性检查维度，至少包含一次用户修正、一次必要澄清和一次工具返回.

20.9. 举例说明回答字符串匹配为何不能可靠确定监督跨度，并给出结构化替代方案.

20.10. 构造两组方向相同和相反的任务梯度，用式（20.16）比较其局部保持趋势.

20.11. 对具有多个正确回答的问题，分别设计条件似然指标与生成质量指标，说明两者可能不一致的原因.

20.12. 设计一份同时衡量新领域收益和旧能力保持的模型选择规则，说明测试集何时才能使用.

第二十一章　参数高效微调

21.1. 对 $d_i = 4096, d_o = 1024, r = 16$ 的投影计算 LoRA 参数量与比例，说明为什么不能使用方阵公式.

21.2. 从迹微分完整推导 $\nabla_A \mathcal{L}$、$\nabla_B \mathcal{L}$ 与 $\nabla_X \mathcal{L}$，逐步核对张量形状.

21.3. 在本章数值例题中继续执行第二次梯度下降，分别计算两个因子的更新. 解释为什么应同时用更新前的因子求本步梯度.

21.4. 证明对可逆 R，$(BR)(R^{-1}A) = BA$，并讨论这对通道可解释性和因子平均的影响.

21.5. 若 A 固定而只训练 B，增量行空间受到何种限制？与同时训练两个因子相比有何差别?

21.6. 利用第一步更新公式，比较固定 α 与固定 s 时秩变化对初始期望增量的影响，明确初始化方差是否也发生变化.

21.7. 构造两个秩一适配器，使其增量之和秩为二，验证直接相加因子产生交叉项.

21.8. 对 $Y = D(W_q)X + sBAX$ 说明为什么不需要对量化码值求导却仍必须计算基座输入梯度；给出错误切断计算图的后果.

21.9. 令 $L = 24, d = 2048, p = 20$，比较输入软提示与直接逐层键值前缀的参数量，并说明参数量之外还应比较哪些成本.

21.10. 对瓶颈 Adapter 证明：若非线性替换为恒等且偏置固定，可以合成一个仿射映射；进一步说明何时增量才是纯线性的. 保留非线性时为何一般不能如此合并?

21.11. 分析将键向量缩放移动到 softmax 之后是否等价，并用单查询、两个键的例子说明原因.

21.12. 设计角色适配的稳定信息、运行时配置和实时观察三层结构，分别给出数据划分、非触发控制与回滚条件.

21.13. 设 $W \in \mathbb{R}^{d\times d}$ 满秩，LoRA 增量为 BA，其中 $B \in \mathbb{R}^{d\times r}$、$A \in \mathbb{R}^{r\times d}$ 且 $r < d$. 证明增量秩的上界，并给出 $W + BA$ 仍满秩的具体例子. 解释为何增量秩不能直接视为多层模型的能力上限，以及增大 r 为何未必改善验证效果.

第二十二章　训练工程优化

22.1. 冻结基座有 2×10^9 个参数，每参数实际存储 0.6 字节；可训练参数有 2×10^7 个，每参数训练状态 16 字节，激活和运行时占 4×10^9 字节．计算总峰值预算，并说明将可训练参数减半的收益．

22.2. 两个微批次分别有 3 和 9 个有效监督词元，平均梯度分别为 2 和 6．求词元平均梯度，并解释为什么累积步数不能替代标签计数．

22.3. 设 $L=100$、$A=20$ MiB，求均匀重计算模型的最优区间及激活存储，并列出实际峰值中还需计入的项．

22.4. 一项适配器微调在增加序列长度后显存不足．设计完整更新步的显存测量与消融方案，区分基座权重、可训练状态、激活和临时工作空间．说明何时选择微批量调整、重计算或状态分片，以及普通数据并行为何可能无法解决问题．

第二十三章　预训练原理

23.1. 对三个有效目标概率 $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$，计算平均负对数似然与困惑度，并说明几何平均的含义.

23.2. 两个微批次分别有 3 和 9 个有效目标，写出其梯度在一次更新中的权重. 若误用等权平均，短批次的词元相对被放大多少倍?

23.3. 为两篇独立文档各含 BOS、三个正文词元与 EOS 的装箱块，列出全部有效预测配对，并给出注意力隔离规则.

23.4. 两个来源按文档等概率抽样，平均有效目标数分别为 256 和 1024. 求实际词元比例，并求词元等权所需文档抽样概率.

23.5. 根据式（23.13）推导 $m = 2n$ 时不同对象覆盖率的大样本近似，解释为什么它不等于两遍无重复遍历.

23.6. 对 $\alpha = \frac{1}{3}, \beta = \frac{1}{2}$，求 N_*、D_* 和最优损失余量随计算预算变化的指数.

23.7. 从拉格朗日条件解释：为什么最优点一般不满足两个损失余量相等?

23.8. 将数值例中的计算预算扩大九倍，求连续最优参数、词元量和损失；再加入参数不超过两十亿的约束.

23.9. 给出一个掩码恢复任务和一个自回归任务，说明为什么两者的平均交叉熵不能直接用于判断预训练方案优劣.

23.10. 设计一组能分别识别参数受限与数据受限趋势的实验点，说明同一运行的多个检查点为什么不能替代独立规模实验.

23.11. 若长期部署有严格参数上限，讨论训练计算最优配置与生命周期最优配置可能出现的差异.

23.12. 说明数据游标、混合日程和监督计数缺失时，权重加载成功为何仍不足以证明正确恢复.

23.13. 领域语料规模固定，但训练预算可以增加．设计一组受控比较，考察增加重复轮数、混合通用数据、扩大模型和提前停止的效果．明确应记录的独立数据量与总训练词元量，并给出支持继续训练或停止的证据．

第二十四章　强化学习基础

24.1. 为一个“搜索—读取—回答”的工具任务定义状态、观测与动作，说明哪些环境信息不可见.

24.2. 将两步算例改为 $\gamma = 0.9$，重新计算 $V(s_0)$ 和两个参数梯度，明确外部折扣因子.

24.3. 从回报递推推导 Bellman 方程，并说明 $\gamma = 1$ 时压缩映射证明失效的原因.

24.4. 用三条轨迹枚举证明两步算例中 ψ 的得分函数梯度与直接求导一致.

24.5. 证明动作无关基线不改变梯度期望，并构造一个动作相关基线使该结论失效.

24.6. 推导包含自身样本的均值基线所产生的 $(1 - \dfrac{1}{K})$ 因子，并说明 $K = 1$ 的情况.

24.7. 将 GAE 数例的 λ 分别改为 0 和 1，计算两步优势并解释差异.

24.8. 一条任务仅因批次切片结束而未终止，说明价值自举与优势递推应怎样处理边界.

24.9. 连续十步概率比均为 1.1，计算轨迹权重并讨论更长序列的方差风险.

24.10. 给出正、负优势各两个概率比，判断 PPO 哪一项生效；解释为什么裁剪不等于硬概率约束.

24.11. 对相同奖励、不同回答长度，比较原始序列回报、$\dfrac{R}{T}$、折扣回报与每步成本四种目标.

24.12. 说明为什么参考策略概率不能替代采样行为概率，以及硬词表截断如何影响支持条件.

第二十五章　人类反馈学习

25.1. 推导二元奖励模型损失对两个奖励输出的梯度，解释它们为何符号相反.

25.2. 构造比较图不连通的例子，说明哪些奖励偏移无法由数据识别.

25.3. 由式（25.6）推导式（25.7），说明有限支持假设的作用.

25.4. 将两回答算例的 β 改为 2，计算最优策略与目标值.

25.5. 证明同提示奖励加常数不改变最优策略；讨论不同提示的常数是否影响跨提示奖励直接比较.

25.6. 从 DPO 损失推导梯度，并比较间隔为零、很正和很负时的系数.

25.7. 为三个回答构造优选绝对概率下降而偏好间隔上升的另一例子.

25.8. 解释序列平均对数概率为何一般不是原序列分布的对数概率.

25.9. 推导序列 KL 的链式表达，指出采样分布改变时哪一步不再成立.

25.10. 比较组内总体标准差与样本标准差对同一组优势幅度的影响.

25.11. 设计参考概率缓存的身份字段，分析模板或终止标记变化后的失效条件.

25.12. 针对“回答越来越长、奖励上升、事实性下降”的现象，提出可区分标注偏差、评分漏洞和候选分布漂移的评估设计.

第二十六章　可验证推理训练

26.1. 为算术等价、代码单元测试和数据库工具任务分别定义“通过”的范围，说明各自仍无法证明什么.

26.2. 设基线正确率为 0.02，正确通过率为 0.95，希望筛选后正确率至少 0.9. 求错误通过率的最大允许值.

26.3. 对奖励 $(0,0,0,1)$，分别使用分母 G 与 $G-1$ 计算标准化优势. 再计算留一基线对应的未标准化优势，解释三种结果的区别.

26.4. 在二元奖励 $p=0.05$ 时，求 $G=2,8,32$ 的混合奖励组概率表达式. 如何把该指标与生成成本联合比较?

26.5. 重现式（26.11），说明哪些步骤使用了条件独立性. 若候选通过共享搜索树产生，原推导还能否直接成立?

26.6. 对两词元回答，概率比分别为 1.5 与 0.5，求序列比. 构造另一个平均词元比相同但序列比不同的回答.

26.7. 在完整组内例题中，将裁剪宽度改为 0.1，重新计算目标和每个位置的局部梯度，明确处于边界时采用的次梯度约定.

26.8. 从式（26.21）证明非负性及当前策略采样下的期望. 说明把行为样本当作当前样本为何会改变估计.

26.9. 比较按回答、按批量词元及固定分母归约. 对相同负优势的长度 2 与长度 20 回答，分析每词元权重及其可能引入的训练行为.

26.10. 对 $n=20, c=4$，计算 $k=1,2,5$ 的 pass@k 组合估计. 说明为什么该指标不等于多数投票返回准确率.

26.11. 若每题在第一次通过后停止生成，设计直接评估该服务策略的记录方式. 说明为何不能无条件把变长样本数代入固定样本组合估计并宣称无偏.

26.12. 一个验证器奖励所有“工具执行成功”的轨迹. 设计一种完成了操作但未完成任务的反例，并提出可核验的终点状态条件.

26.13. 建立按题族隔离的推理蒸馏数据流程，分别指出教师、验证器调参和自适应测试可能引入污染的位置.

26.14. 训练奖励升高、回答变长、独立正确率不变. 提出可区分长度偏置、验证器利用和问题分布变化的证据方案.

第二十七章　模型规模及资源预算

27.1. 取消 $h_q d_h = d$ 假设，重新推导注意力参数量，并说明式（27.2）第二行何时失效.

27.2. 将参数例题改为独立输出矩阵、每层两组及末端归一化均改为带偏置的 LayerNorm，其他投影仍无偏置，重新计数全部参数.

27.3. 对矩阵乘法完整写出前向与两条反向的形状和操作量，说明冻结权重后哪些梯度可能仍需计算.

27.4. 对长度分别为 1000、2000、3000 的请求及页长 256，比较逻辑缓存位置数与独立分页后的保留位置数.

27.5. 某量化方案每 128 个参数保存一个四字节尺度，权重四位. 求每参数平均存储字节数，并解释为什么还可能低估实际权重文件.

27.6. 在 16 字节状态配置下将梯度改为四字节，重新推导 DDP 与各分片阶段的状态量.

27.7. 联合预算例题中，利用率降至 .3 且期限仍为 10 天，求计算约束与八设备拓扑取整结果.

27.8. 检查点每 20 分钟产生 400 GB，后台有效写带宽为 200 MB/s. 队列能否长期稳定？给出同一单位下的推导.

27.9. 设计包含长度峰值、数据供给降速和通信效率下降的三个压力情景，说明哪些影响已被利用率吸收，哪些必须单独加项.

第二十八章　算力基础设施

28.1. 某 SM 具有 65536 个寄存器、64 KiB 共享存储、2048 线程上限与 16 块上限. 每块 256 线程，每线程 64 寄存器，每块 24 KiB 共享存储. 忽略分配粒度，求驻留块数上界；将共享存储改为 16 KiB 后，哪个约束成为主导?

28.2. 对 $X \in \mathbb{R}^{B\times 4096}$、$W \in \mathbb{R}^{4096\times 4096}$，每元素两字节，求理想算术强度达到 50 FLOP/byte 所需的最小整数 B. 解释为何这不足以保证达到峰值算力.

28.3. 一项算子包含 4×10^{12} FLOP，跨 HBM 传输 10^{11} 字节. 峰值为 2×10^{14} FLOP/s，带宽为 2×10^{12} byte/s. 求时间下界，并比较操作量减半与流量减半的效果.

28.4. CPU 绑定到 NUMA 0，而输入页已经在 NUMA 1 分配，GPU 位于 NUMA 0. 说明仅更改线程亲和性为何不足，并画出可能的数据路径.

28.5. 读取、整理、传输和计算分别需要 10、20、8、16 毫秒. 求串行批次时间与理想稳定流水间隔；若整理与读取共享瓶颈资源，说明原流水估计应怎样修正.

28.6. 使用式 (28.12)，比较 $p=8$ 与 $p=64$ 在相同消息和链路参数下的启动项与传输项. 为什么全局设备数增加不代表一次集合通信变快?

28.7. 给出一个端口总带宽很高、跨割带宽却很低的交换拓扑. 说明专家全互换为什么可能比局部邻近通信更容易暴露此问题.

28.8. 一个存储路径支持 1 GiB/s，随机请求能力为每秒 5000 次. 分别计算 4 KiB 与 1 MiB 请求的理论有效带宽上界，并说明增加并发为何不能无限提高吞吐.

28.9. 将本章检查点写入窗口从 120 秒改为 90 秒，重新计算集群及每节点所需带宽. 若允许分层暂存，分别指出前台暂停和远端持久化的完成条件.

28.10. 为跨 64 个分片的检查点设计提交清单，说明如何处理第 63 个分片完成后进程失效，以及恢复时发现一个分片摘要不匹配的情形.

28.11. 从式（28.23）推导最优间隔. 若作业平均故障间隔缩短到原来的四分之一，最优间隔怎样变化？哪些实际条件会使该近似失效？

28.12. 某训练作业设备忙碌率提高，但单位有效词元能耗也提高. 至少给出三种机制，并说明需补充哪些测量才能区分它们.

28.13. 为一个国产加速器迁移任务列出兼容矩阵的最小字段，明确“模型加载成功”“算子正确”“通信可用”“端到端达标”各自能够支持什么结论.

第二十九章　分布式训练

29.1. 构造三副本、每副本两个不等长微批次的计数和梯度，比较三种归一化：微批均值平均、副本均值平均、全局词元平均.

29.2. 在采用平均归约的通信组中，证明式 (29.4)；若梯度已被框架额外除以累积步数，应如何修正？

29.3. 用反例说明先裁剪局部梯度再归约不等于先归约再裁剪.

29.4. 在重计算近似模型中令 $L = 100$、每激活单位 20 MiB，求最优整数区间及近似存储，并列出未计入的内存项.

29.5. 推导环式 Reduce-Scatter 的轮次和每成员发送量，再说明如何与 All-Gather 组成式 (29.6).

29.6. 对状态算例把分片组改为 4，重新计算各阶段持久状态并讨论峰值可行性.

29.7. 为张量并行例题计算各权重片梯度，并验证拼接后与完整矩阵微分一致.

29.8. 证明非线性一般不能跨越部分和归约，指出式 (29.12) 为什么允许本地执行非线性.

29.9. 流水阶段数为 8 时，需要多少微批次才能使理想气泡比例不超过 10%？讨论该选择的显存代价.

29.10. 由两个块的 Softmax 合并式推导三个块的组合，并解释全掩码块的处理.

29.11. 给出 64 设备的合法逻辑网格，明确不同通信组与有效词元计数范围.

29.12. 设计改变数据并行度后的续训声明，区分资产完整、目标一致、数值接近和逐位复现.

第三十章　增量解码

30.1. 提示长度为 5、输出长度为 4，列出每次前向处理的词元位置、产生的预测与最终缓存长度.

30.2. 逐层证明旧位置隐状态不变，并指出双向注意力在哪一步破坏该证明.

30.3. 改变数值例的新键为 $\log 4$，重新计算权重与输出；说明漏掉自身键值的错误.

30.4. 对提示 8、输出 4，分别计算全量重算与缓存方案的线性处理位置数和因果合法分数对数.

30.5. 设 40 层、16 个 KV 头、键维 128、值维 64，元素均为 2 字节，缓存 16384 词元，计算裸容量.

30.6. 保持 32 个查询头，比较 4 个 KV 头与 1 个 KV 头的容量；解释为什么计算不能简单按容量同比减少.

30.7. 推导式（30.10），并用维度说明 U_K 不能与任意位置旋转交换.

30.8. 块大小 8、长度 19、页表 $(4, 1, 7)$，求位置 17 的物理块和偏移及尾部空位.

30.9. 四条请求共享 16 词元前缀、各有 5 词元后缀、块大小 8，计算共享前后块数与尾部碎片.

30.10. 说明未满共享块追加时为何需要写时复制，以及引用计数归零为何仍可能暂不可回收.

30.11. 区分全注意力历史淘汰、原生滑窗回收、无引用前缀淘汰和活动请求抢占.

30.12. 设计缓存身份字段，分别讨论仅改变采样温度、适配器、模板和图像输入时是否可复用.

第三十一章　批处理调度

31.1. 某请求在 0 秒到达，0.1 秒开始执行，0.3 秒收到首词元，0.9 秒收到第 31 个词元．计算 TTFT、TPOT 和首执行等待，并指出哪些量可以相加．

31.2. 将图（31.1）的轮时改为单序列 6 毫秒、双序列 10 毫秒，重新计算两种策略的完成时间．解释固定轮时假设的影响．

31.3. 对长度为 64、64、256 的三个提示，计算线性填充与完整平方注意力的长度利用率．为何不能将结果直接用于变长内核？

31.4. 从逐查询可见键数推导式（31.4），并解释分块后注意力数学总量与历史重读次数的区别．

31.5. 在缓存算例中，已有 257、255 长度的序列各再生成两个词元，需要多少页？说明页内碎片为何会影响边际接纳成本．

31.6. 给出一个系统人数的阶梯函数，直接计算驻留面积以验证式（31.7）．解释有限观察窗两端截断如何处理．

31.7. 推导 M/M/1 的空闲概率、平均队列长度和式（31.13）．指出连续批处理违反其中哪些假设．

31.8. 当两个随机阶段的 p99 分别为 100 与 200 毫秒时，能否断言总时延 p99 为 300 毫秒？使用并集界给出一个成立的陈述．

31.9. 构造每请求公平与每词元公平结果相反的两租户负载，并说明长上下文为何进一步影响成本公平．

31.10. 输出产生速率为每秒 12 KiB，客户端消费每秒 4 KiB，空缓冲容量为 64 KiB．求填满时间，并设计到达上限后的状态转换．

31.11. 设计一个记录取消、页释放和重新入队的事件序列，证明调度器不会让已终止请求复活，也不会重复释放共享页．

31.12. 比较开放环和闭合环压测，列出一个能识别过载、重算放大和慢消费者的证据方案，并明确所有统计分母．

31.13. 服务显存尚有余量，但尾延迟持续恶化. 设计一次诊断，区分持续超载、长提示阻塞解码与慢客户端造成的输出积压；说明不同证据下应调整准入、调度还是扩容，并给出扩容生效前的处置原则.

第三十二章　模型服务架构

32.1. 同一权重修改模板后是否仍是同一服务修订？分别从模型计算与接口语义解释.

32.2. 为取消与完成同时发生的请求设计状态更新规则，并说明资源释放为何还需要独立幂等保护.

32.3. 一条流已输出十个事件后连接中断，列出支持恢复所需的状态；说明事件重放与重新生成的区别.

32.4. 请求总预算 2 秒，排队已耗费 1.2 秒、网络与预处理耗费 0.3 秒，计算剩余预算并讨论是否应再发起一次 1 秒重试.

32.5. 区分启动、存活与就绪信号，说明过载时错误使用存活重启会发生什么.

32.6. 对 2 GiB 缓存、40 GiB/s 有效带宽，计算传输下界；每秒 30 请求时总带宽需求是多少?

32.7. 在 P/D 迁移中，接收端已获得数据而所有权提交结果未知，应怎样避免双重输出?

32.8. 将容量算例启动时间改为 120 秒，计算新增积压，并提出两种不依赖即时扩容的处理方式.

32.9. 比较影子、金丝雀与蓝绿方式的资源和证据边界，解释影子工具副作用为何需要隔离.

32.10. 两个副本的 P99 分别为 1 秒和 5 秒，为什么不能直接报告全局 P99 为 3 秒?

32.11. 失败概率 0.2、最多四次总尝试，计算期望重试放大系数，并说明独立假设在过载时为何可疑.

32.12. 制定一次工作进程失联演练的观察项，覆盖在途输出、缓存释放、错误计量和剩余容量.

第三十三章　推理计算优化

33.1. 推导在线统计量合并的三块表达，并说明精确算术下合并顺序无关的原因.

33.2. 为两个输出不同且指数质量不等的键块构造数例，比较正确合并与局部等权平均.

33.3. 说明全掩码块与全掩码行的区别，并设计不产生无穷相减的初始化规则.

33.4. 从 Softmax Jacobian 推导本章反向公式的行内减均值项.

33.5. 某算子占总延迟 20%，加速 10 倍，求端到端理论加速上限.

33.6. 三个形状桶的编译成本分别为 2、10、1 秒，预计调用次数分别为 1000、100、20 次，单次节省分别为 3、50、10 毫秒. 忽略缓存维护等额外成本，判断哪些桶值得预编译.

33.7. 用 $p=(0.1,0.6,0.3)$、$q=(0.5,0.2,0.3)$ 计算接受率、残差及最终输出质量.

33.8. 证明接受率等于 1 减去总变差距离，并讨论草稿与目标支持集不相同的情况.

33.9. 构造温度不同的草稿和目标，说明精确采样为何不要求温度相同，却要求使用实际概率.

33.10. 候选条件接受率依次为 0.9、0.7、0.5，求每轮期望提交词元数.

33.11. 描述第二个候选拒绝后目标 KV、草稿 KV、已输出长度和未物化后缀应如何更新.

33.12. 推导窗口截断的输出误差上界，解释它为何不直接提供下游任务质量保证.

33.13. 同一负载下，方案 A 每轮草稿生成、目标验证及同步共耗时 12 ms，平均提交 3 个词元；方案 B 的接受率更高，但每轮耗时 20 ms，平均提交 4 个词元. 按长期平均成本判断应选哪一方案，并说明迁移到高并发服务时还应检查哪些条件.

第三十四章　模型优化

34.1. 对目标范围 $[-0.7, 1.8]$ 与无符号三比特码计算初始尺度、取整零点及实际恢复端点，说明端点为何发生偏移.

34.2. 推导均匀误差模型的 $\frac{s^2}{12}$，并构造一个全部输入恰在格点上的反例，说明其为何不是恒等式.

34.3. 比较取偶数最近舍入与向零截断在正负半格点附近的误差. 为一个可复现制品写出完整舍入约定.

34.4. 对本章异尺度矩阵计算逐张量和逐行的整体均方误差，再解释为何误差改善不能直接推出任务改善.

34.5. 给定 4096×4096 权重、四比特码、每行每 128 个元素一组、每组四字节元数据，计算忽略对齐的存储与相对十六位基线压缩率.

34.6. 从约束二次目标推导式（34.14），说明阻尼改变了什么，以及输入完全不相关时为何不发生跨坐标补偿.

34.7. 若 GPTQ 先处理另一坐标，离散结果是否必然相同？用两维例子或明确论证说明处理顺序的作用.

34.8. 用式（34.17）说明 AWQ 与 SmoothQuant 的共同基础，再分别写出二者关注的误差对象.

34.9. 给定激活通道范围 $(64, 4)$ 与权重列范围 $(1, 4)$，计算 $\alpha = \frac{1}{2}$ 的 SmoothQuant 尺度及两侧新范围.

34.10. 证明均匀量化函数的真实导数不能直接提供一般的有效梯度，并说明 STE 的近似性质与部署匹配要求.

34.11. 比较 E4M3 与 E5M2 的指数和尾数预算，说明为何最大值、下溢和累加行为必须按具体格式与内核声明.

34.12. 从式（34.22）推导对称权重量化可省略的修正项，并讨论分组尺度为何改变累加组织.

34.13. 构造一个模型权重减少四倍但请求时延改善很小的负载，按带宽、计算、缓存和其他时间解释原因.

34.14. 对键值量化的一阶误差式说明：为何只测值缓存的逐元素均方误差不能覆盖注意力分配变化?

34.15. 证明在保留数值不变、仅固定非零数量的条件下，幅值剪枝最小化权重平方误差；给出不最小化任务损失的反例.

34.16. 使用式（34.27）分析非驻点时一阶项的作用. 为什么接近训练结束也不能未经测量就认定所有 $g_q = 0$?

34.17. 对 $H = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$、$w = (1, 1)^{\mathsf{T}}$，计算分别删除两个权重时的二阶补偿与代价.

34.18. 写出门控前馈层删除中间通道的三个矩阵索引关系，并说明偏置和优化器状态应如何同步.

34.19. 两字节数值配两字节索引时，忽略行指针求稀疏存储低于稠密存储的非零比例阈值. 为什么满足它仍不保证加速?

34.20. 对奇异值 $(10, 4, 3, 1)$，计算秩 1、2、3 近似的绝对与相对 Frobenius 误差，讨论给定参数预算能否实现某个误差目标.

34.21. 从列空间投影出发复现截断 SVD 的最优性证明，指出截断奇异值重合时唯一性为何可能失效.

34.22. 对 $m = 4096, n = 11008$，求减少因子参数所需的秩阈值. 分别计算 $r = 512$ 与 $r = 2048$ 的参数比例.

34.23. 对 $W = \mathrm{diag}(4, 1)$、$C_x = \mathrm{diag}(1, 25)$，比较普通秩 1 SVD 与输入加权近似的输出误差.

34.24. 推导温度平方缩放后的学生 logit 梯度，并说明高温近似需要中心化 logits 的原因.

34.25. 在蒸馏数例中将 λ 改为 0.8，求损失与学生梯度；若硬标签改为第二类，两项监督如何竞争?

34.26. 教师与学生使用不同分词器，列出逐词元 KL 不成立的两个原因，并解释序列蒸馏保留和丢失了哪些信息.

34.27. 对因子变换 $A' = 10A, B' = \dfrac{B}{10}$，说明乘积不变而量化误差可能变化的原因.

34.28. 某方案报告权重文件缩小四倍、单层 FLOPs 减少两倍. 设计能够确认端到端实际收益的最小证据集合，并明确不能从已有两个数字推出什么.

34.29. 剪枝、低秩分解和量化分别作用于原始模型时均达到质量要求. 说明这些结果能否证明顺序组合后的模型也达标，并设计同时覆盖累计误差、处理顺序与实际部署收益的验收方案.

第三十五章　双向语言建模

35.1. 给定一条句对输入，分别写出词元、位置和句段编号，说明句段不同为什么不自动阻断注意力.

35.2. 在掩码数值例题中，将第 1 个键改为填充，计算第 2 个查询在双向与因果条件下的权重和输出.

35.3. 若从 20 个候选位置独立以 .15 概率选择监督，求空集合概率. 若改成固定选择三个位置，讨论选择指示变量是否独立.

35.4. 假设随机替换从包含原值的 V 个词元中均匀选择，计算经典三分支配方下被选中位置实际保持原值的概率，并区分它与进入保留分支的概率.

35.5. 两条序列分别有一个和三个监督词元，平均损失分别为 .2 与 .8. 计算逐序列平均和逐监督词元平均，解释训练目标的差异.

35.6. 推导式 (35.11) 中共享词元表的输出路径梯度，并说明为什么还要加入损坏输入查表路径的梯度.

35.7. 构造一组起止分数，使分别取最大值产生终点早于起点的结果；在最大答案长度为二的约束下求最优合法片段.

35.8. 使用本章伪似然例题验证总质量为 1.64. 若将结果另行归一化，为什么仍不能宣称这些条件预测就是新联合分布的条件概率?

35.9. 为同一原文的两套损坏策略设计比较方案，区分参数质量变化与任务难度变化，并说明验证位置是否需要固定.

第三十六章　混合专家模型

36.1. 对 $E_r = 4, k = 2$、分数 $(\log 0.4, \log 0.3, \log 0.2, \log 0.1)$，计算选中内归一化和保留全专家概率的两种权重. 设已选专家输出为 $(1, 0)$ 与 $(0, 2)$，比较最终输出.

36.2. 从式（36.8）证明：固定选中集合且 $k = 1$ 时，选中内归一化的路由任务梯度为零. 再解释保留全专家概率为何不必为零.

36.3. 对某个词元，两个专家输出相同. 分析选中内归一化时路由权重的任务梯度，并说明这是否意味着专家参数梯度也为零.

36.4. 构造 $E_r C \geqslant Nk$ 但仍存在容量溢出的例子. 分别计算路径拒绝率、受影响词元比例和完全失去专家处理的比例.

36.5. 在本章数值例题中，把专家 2 接收规则改为按该专家的路由权重降序选择，并列时按词元编号升序接收. 确定哪条路径被拒绝，再比较输出变化. 说明规则不同为何不能只当成排序实现细节.

36.6. 推导式（36.17）. 如果辅助统计覆盖两个大小不同的本地批次，写出正确的全局平均概率，解释局部均值直接平均的问题.

36.7. 令 $d = 512, d_f = 2048, E_r = 16, k = 2$，忽略偏置，计算门控专家的总参数与每词元激活参数. 加入两个始终执行的同尺寸共享专家后，两个计数分别如何变化?

36.8. 保持每词元专家矩阵计算近似不变，将专家中间宽度减半、专家数与选择数加倍. 哪些成本不能随专家矩阵算术量保持不变?

36.9. 对 Soft MoE 分发与合并矩阵，分别验证列和与行和. 举例说明若先混合全序列词元再调用专家，为什么早期输出可能读取未来输入.

36.10. 某模型在训练大批次中专家利用率较高，在低并发解码中却很慢. 用矩阵形状、权重读取、通信延迟和路由负载解释至少三种可能机制.

36.11. 某报告只给出平均专家负载和激活参数量，便断言单卡内存与吞吐满足上线需求. 说明尚缺哪些可支持结论的结构信息和运行证据.

第三十七章　长上下文建模

37.1. 将 $T_0 = 4096$ 扩展为 $T_1 = 32768$，分别计算采用长度比与端点比的位置映射，比较最后一个位置的映射结果．说明为什么配置必须统一这两种约定．

37.2. 对给定频率 $\omega = \dfrac{1}{256}$ 和距离 $\Delta = 2048$，计算扩展因子为四时的插值相位，并写出固定查询与键条件下的点积分量变化上界．

37.3. 从式（37.11）推导频率比例，验证最高与最低频带的边界．解释为何 $d_r = 2$ 不适用该基数规则．

37.4. 对式（37.3），给定目标权重 0.9，推导序列长度由 101 增至 901 时相关分数所需增加的量．说明真实多头网络与该构造相比多出了哪些变量．

37.5. 因果窗口包含最近八个位置，网络有六层．求理论最大连续可达范围，并画出从最早位置到最终位置的一条路径．解释可达与可用的区别．

37.6. 在仅解码器模型的开头加入一个“全局词元”，是否能让它在同一次因果前向中汇总整个后续文档？利用图上的信息方向证明结论．

37.7. 两个长度桶分别为 2048 和 16384，要求实际训练词元占比相等，求按序列抽样的概率．再说明样本重复率为何仍需单独记录．

37.8. 使用本章 KV 参数，计算上下文 65536、生成上限 8192、并发三条时的裸缓存容量，并列出尚未包含的显存项．

37.9. 设计一个长度、证据位置和干扰程度彼此独立的多证据问题，给出正确答案、无答案变体与评分规则．说明仅测试最长长度的开头位置会遗漏什么．

37.10. 对需要逐条核对长合同数字的任务，比较摘要压缩、检索和直接长窗口三种方法．指出每种方法可能丢失的信息，以及应保留的来源与中间状态．

第三十八章　开源模型体系

38.1. 推导式（38.1）在行向量约定下的形式，并核对查询吸收与值求和后的上投影形状.

38.2. 对给定矩阵 U 和两个位置旋转 R_1, R_2，说明为什么一般不能为所有历史位置复用一个 $U^{\mathsf{T}} R_j^{\mathsf{T}} R_t q$. 给出可以交换的特殊条件.

38.3. 将本章缓存例题的键值头数由四改成一，比较 MQA 与潜在缓存，说明压缩方法之间不存在脱离维度的固定优劣.

38.4. 一个 MoE 有总参数 40 十亿、每词元激活 4 十亿，采用每参数两字节. 若不分片或卸载，为什么不能把权重驻留估成 8 十亿字节?

38.5. 同一权重使用两种不同对话模板，输出质量不同. 如何区分模型能力差异与输入协议差异?

38.6. 对混合状态缓存公式分别令 $L_a = 0$ 和 $L_s = 0$，解释两个极端. 再说明固定状态大小为何不保证无限长依赖恢复能力.

38.7. 一个模型卡声明词表大小为 V，嵌入有 $V + 16$ 行. 列出需要补充的证据，不直接作正确或错误判断.

38.8. 设计一个比较推理模式与直接回答模式的任务表，明确正确率、总生成长度、工具轮数、时延和重试预算.

38.9. 选择一个未在本章展开的开放权重检查点，按结构、训练、协议、模态、许可和制品六维建立来源清单；将公开声明与自己尚未验证的项目分开.

38.10. 某模型同时使用多头潜在注意力（MLA）和混合专家（MoE）. 分别指出二者改变的计算模块，并为长上下文解码与专家分布式执行列出需要单独核算的状态、计算和通信. 说明为何每词元激活参数量不能代表全部部署显存.

第三十九章　视觉 Transformer

39.1. 对 $B=2, C=3, H=96, W=128, P_h=8, P_w=16$，计算 G_h, G_w, N, F；写出最后一个补丁和块内最后一个像素的索引.

39.2. 证明式（39.2）可逆，并说明当滑窗重叠、忽略边界或带填充时，哪些结论发生变化.

39.3. 以本章 4×4 图像算例为基础交换块内两个像素的展平位置. 怎样变换 W_E 才保持输出不变？只改变展平规则会造成什么错误？

39.4. 推导重叠滑窗情况下图像梯度的累加公式，并说明为什么不能用无重叠逆重排直接替代.

39.5. 当 $F=12, d=5$ 时，证明至少存在七维的补丁扰动空间不会改变线性嵌入. 这个结论是否足以判断分类性能？

39.6. 证明没有位置编码时，固定 CLS、任意置换补丁不会改变分类读出. 加入位置后同时置换内容和位置，与只置换内容有何区别？

39.7. 把 2×2 标量位置网格插值至 4×3，明确角点或中心对齐规则，并解释为什么不能将 CLS 当作额外空间点.

39.8. 当 $d=512, \rho=4$ 时，求式（39.21）中线性项与二次项相等的序列长度；解释这为何不是硬件延迟交点.

39.9. 比较固定补丁尺寸提高分辨率与固定分辨率减小补丁尺寸：哪些参数形状变化，哪些只是激活形状变化？

39.10. 写出带填充掩码的 GAP 读出梯度，并说明直接除以包含填充的 N 会产生什么偏差.

39.11. 为包含小字符的图像分类任务分析随机裁剪、镜像与 Mixup 的标签假设. 提出一个保持划分固定的比较方案，区分增强收益与数据泄漏.

39.12. 一个模型张量形状全部正确，但部署准确率骤降. 按输入范围、通道、位置映射、读出和标签映射构造逐层排查顺序，并为每一步指出所需证据.

第四十章　多模态建模

40.1. 推导不重叠块投影与卷积等价所需的权重排列，说明增加补齐、重叠或不同步幅后哪些关系改变.

40.2. 一幅图像两边都放大 1.5 倍，块大小不变且尺寸仍可整除，推导词元数和完整自注意力位置对数的比例.

40.3. 从按行、按列交叉熵分别推导式（40.7），说明 R 和 C 一般不相同.

40.4. 将双样本例题温度改为 0.5，重新计算损失、逻辑值梯度和对数缩放梯度；比较温度对概率和梯度链的影响.

40.5. 推导向量归一化雅可比，验证其输出与单位向量正交，说明零向量处为何需要另行约定.

40.6. 构造一图多文本的正对集合，比较均匀多正对交叉熵与正对总概率损失的梯度方向.

40.7. 在重采样例题中将查询改为 0，计算输出及值梯度，并解释此时为何表现为平均但一般重采样并非平均.

40.8. 对 $M < N_v$ 的固定注意力矩阵构造一个非零零空间扰动，说明固定映射的信息损失与非线性重采样器整体能力的区别.

40.9. 根据表（40.2）写出输出逻辑值与目标的三个配对，解释把提示位置输出全部删除会丢失哪一个答案监督.

40.10. 语言主干被冻结、连接器可训练时，画出必要梯度路径；说明在语言主干周围整体停止自动微分为何会改变训练.

40.11. 对 3 秒、16000 样本/秒的音频，采用 400 样本窗口与 160 样本步长且不补齐，计算窗口数，再说明中心补齐为何改变结果.

40.12. 构造持续 0.2 秒的事件，在 0.8 秒均匀采样间隔和随机相位假设下计算命中概率，并说明它与视觉识别正确率的关系.

40.13. 用有偏移和漂移的两台设备设计时间戳校正，给出窗口代表时间和实际可用时刻的区别.

40.14. 为视觉计数或声画匹配任务设计受控样本对、缺失模态对照和任务指标，说明仅验证输出变化为何不充分.

第四十一章　扩散模型的概率原理

41.1. 用归纳法和望远镜求和分别证明前向条件方差为 $1-\bar{\alpha}_t$.

41.2. 推导原数据均值不为零、协方差不为单位阵时 x_t 的无条件均值与协方差.

41.3. 固定同一总噪声画出各时间状态，与独立单步噪声构成的路径有何联合分布差异？解释为何二者每个时刻的边缘仍可一致.

41.4. 完整展开高斯指数并配方，推导后验均值和方差；说明 $t=1$ 的退化为何不能直接进入高斯 KL.

41.5. 推导式（41.7），分析数据尺度与维数对终端先验误差的影响.

41.6. 从 Jensen 不等式推导负对数似然上界，再将前向链反向分解，得到三个变分项.

41.7. 对固定 $\sigma_t^2=\beta_t$ 推导噪声损失权重；再说明预测方差时哪些 KL 项不能忽略.

41.8. 设计非均匀时间抽样分布，写出内部变分项的无偏估计；区别时间加权与每元素归一化.

41.9. 证明平方损失最优噪声预测是条件期望；为什么这不意味着准确恢复每次加入的独立噪声?

41.10. 验证 x_0、噪声与速度参数化之间的正交变换，并推导各自等权损失隐含的时间权重.

41.11. 由式（41.34）计算 $\eta=0$ 和 $\eta=1$ 的跨步条件方差，解释确定性路径仍可生成分布.

41.12. 修改一维算例的噪声预测误差为 0.2，比较高斯 KL 的变化；说明为什么只用两步的构造不能证明样本生成质量.

第四十二章　条件视觉生成

42.1. 写出视觉与文本交叉注意力的全部形状，比较它与联合注意力的配对数.

42.2. 潜图补丁边长从 4 改为 2 时，位置数与完整打分矩阵如何变化？哪些成本不遵循平方变化?

42.3. 从高斯扰动密度推导条件得分与噪声条件均值的关系，指出 $b_t = 0$ 时的问题.

42.4. 将 $(1+s)\epsilon_c - s\epsilon_u$ 转换成本章引导约定，并计算 $s = 3$ 对应的 w.

42.5. 为什么固定时间的幂次密度解释不能直接证明 CFG 终点分布?

42.6. 推导对角高斯到标准正态的 KL，解释重参数化如何支持梯度传播.

42.7. 计算 $1024 \times 1024 \times 3$ 到 $128 \times 128 \times 16$ 的元素压缩比，解释为何不是 64 倍.

42.8. 证明平方速度回归的条件均值最优性，并写出连续性方程的弱形式.

42.9. 对 $u(z) = z$ 使用四步 Euler，比较结果与 e，解释误差来源.

42.10. 比较完整时空注意力与分解注意力的连接关系，说明为何不能只比较 FLOPs.

42.11. 为图像编辑分别定义条件遵循与未编辑区域保持指标，构造二者冲突的情形.

42.12. 设计生成制品清单，列出替换 VAE、文本编码器或求解器时必须核对的契约.

42.13. 计划把生成器的求解步数由 30 步降至 1 步. 分别说明仅改变数值求解步数与使用适配的一致性训练或蒸馏模型所需的论据，并设计覆盖质量、多样性、条件约束和真实延迟的比较.

第二部分

实验指导

实验总则

实验的目标是验证一个明确机制．开始前应记录问题、输入、对照条件和可证伪判据．若只能运行原理计算部分，应如实界定范围；不要把未运行的模型迁移、GPU 内核或多机部署写为通过．

环境及资源

为实验使用独立的 Python 环境，按 Notebook 的依赖单元准备库并记录实际版本．先检查数据与模型制品的来源、修订、许可及下载规模，再决定是否运行相应单元．CPU 原理计算、GPU 训练和真实服务测试分别计量，不能按 Notebook 文件名假设资源需求．记录设备、可用内存、精度、批次、序列长度和预算；不具备所需设备时，保留该项为未验证．

执行顺序及状态

先阅读整个实验的数据流和资产流，再从初始化单元按依赖顺序执行．修改数据、种子或模型配置后应重新检查下游状态；最终提交前从干净内核重放所声明的执行范围．保存输入制品身份和输出记录，避免隐藏变量使结果只能在当前会话中复现．需要联网下载、较长训练或外部服务的部分应单独列出条件和成本．

数值及性能验证

数值对照使用相同输入、参数、掩码和目标归一化方式；报告最大误差或相对误差，并说明容差依据．随机实验应记录种子和重复次数．性能测量需区分初始化、预热与稳定运行；加速器计时考虑同步，报告输入长度分布与统计量．通过形状检查不意味着数值正确，通过数值对照也不意味着获得性能提升．

提交材料

每份报告至少包含实验目标、环境与资产、操作范围、预期判据、实际证据和失败解释. 图表应可追溯至原始结果；保留错误样本及未通过检查，不只展示成功案例. 若使用构造数据或状态演算，应明确其只能验证算法关系，不能替代真实模型质量、服务吞吐或业务收益.

从理论到实现

先在教材中确定定义与前提，再在本册中完成推导练习，随后在 Notebook 中定位对应对象. 原理实现负责说明机制，生产库负责完整接口与执行路径；迁移时比较的是契约、数值和边界，不能仅因两段代码都能运行便认定它们等价.

框架实践路线

微调框架实践见实验 12，先完成数据、监督位置、训练和适配器重载，再对照集成训练框架. 分布式实践见实验 16，保持训练目标与有效批量，研究状态分片、检查点和恢复. 推理部署实践见实验 17，先完成 vLLM 服务路径，再迁移到 SGLang 并比较质量、负载和故障行为.

三条路线以本册中的命令、配置和配套训练入口为操作材料. 版本号用于定位接口核查基线，完整环境、硬件支持与性能结论须由实际运行建立. 选择框架时先确定职责和验收条件，同类工具的迁移应复用相同数据、模型身份和评价口径.

实验 1　神经网络训练基础

配套文件： `notebook/10_foundations.ipynb`
教材关联： 第 2 章、第 3 章、第 17 章.

实验目标

区分类别编码、概率输出和参数优化，建立可检验的训练基线.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_classification_metrics`
- `my_one_hot`
- `my_sigmoid`

操作及观察

1. 先固定训练、验证划分，检查特征缩放仅在训练集拟合.
2. 对照类别 ID 与 One-hot，逐项核对行和及类别落点；用相同输入比较 NumPy 与 PyTorch 前向输出.
3. 在两层 MLP 上记录输入、隐藏层、logits 和标签形状，再运行训练与验证闭环.
4. 冻结分类阈值后计算混淆结构、精确率和召回率；比较训练损失与留出表现.

结果判读及提交

提交形状表、同输入差异、训练/验证指标及阈值来源. 损失接口接收 logits 时不得预先使用 Sigmoid；训练准确率不能代替泛化证据.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 2　训练数据工程

配套文件：`notebook/20_dataset.ipynb`
教材关联：第 18 章、第 23 章、第 20 章.

实验目标

把记录、Schema、内容身份和划分规则共同视为数据资产.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_locate_project_directory`
- `my_build_parallel_pairs`
- `my_add_quality_columns`

操作及观察

1. 逐条检查候选句对的字段、空值与唯一身份，保留被拒记录的原因.
2. 执行 JSONL 序列化与原子发布，核对重载记录数及内容哈希.
3. 通过 Datasets 加载同一权威源，比较 map/filter 前后的记录保留关系.
4. 生成稳定划分并持久化 Arrow，重载后验证 Schema、样本身份和划分交集.

结果判读及提交

提交数据清单、哈希、划分交集和重载对照. 文件大小相同不证明内容相同；新增语料后须重新审查泄漏，不能只验证行数.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 3　子词分词

配套文件：`notebook/21_tokenizer.ipynb`
教材关联：第 4 章.

实验目标

验证编码可逆性、合并顺序和特殊词元协议.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_locate_project_directory`
- `my_sha256_file`
- `my_is_clean_record`

操作及观察

1. 固定训练文本，跟踪高频字节对合并和词元边界变化.
2. 用已见及未见文本检查编码—解码往返；与 tiktoken 对照时记录词表与预处理差异.
3. 检查 BOS、EOS、PAD 及聊天或翻译模板，构造批次并区分 padding 与截断.
4. 保存并重载分词制品，复查特殊 ID、样本编码和标签屏蔽.

结果判读及提交

提交合并轨迹、往返失败样本、特殊 ID 表和重载结果. 不同训练词表不要求产生相同 ID；目标文本的 padding 不能计入有效损失.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [1]；本实验的运行结果仍须按实际配置单独验证.

实验 4　Transformer 实现

配套文件： `notebook/30_transformer.ipynb`
教材关联： 第 4 章、第 5 章、第 7 章、第 6 章、第 8 章、第 37 章.

实验目标

把注意力、掩码和教师强制连接为可解释的序列到序列实验.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_encode_content`
- `my_encode`
- `my_batch_encode`

操作及观察

1. 核对 ByT5 分词接口、源目标边界、decoder 输入与标签移位.
2. 逐层记录注意力头、前馈层、残差及交叉注意力的形状，验证掩码可见范围.
3. 保持基线划分与预算，训练翻译闭环并从起始词元自回归生成；区分训练损失和生成评价.
4. 使用 Notebook 中的参数映射与契约检查对照 torch.nn.Transformer；分析迁移 MarianMT 时架构与制品的变化.

结果判读及提交

提交因果性检查、移位样本、损失曲线和生成错误分析. 交叉注意力热图只表示内部权重；不能直接称为语言学对齐或因果解释.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [2]；本实验的运行结果仍须按实际配置单独验证.

实验 5　自回归生成

配套文件：`notebook/31_nlp_gpt.ipynb`
教材关联：第 4 章、第 5 章、第 8 章.

实验目标

验证未来信息不可见，以及采样策略改变输出分布的方式.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_encode`
- `my_decode`
- `my_get_batch`

操作及观察

1. 构造上下文与下一词元标签，手工核对首尾位置和有效目标数量.
2. 检查因果注意力矩阵，并比较只改变未来词元时较早位置的输出.
3. 固定模型与提示，比较 temperature、top-k 和 beam search，记录随机种子与预算.
4. 将输入与生成配置映射到 Transformers 接口，核对终止词元与最大生成长度.

结果判读及提交

提交因果性差异、采样配置及生成样本. 单条流畅输出不能证明训练有效；不同解码预算的质量不能直接归因于模型变化.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 6　模型评估

配套文件：`notebook/50_model_evaluation.ipynb`
教材关联：第 14 章、第 9 章.

实验目标

建立从逐例预测到发布决策的证据链.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_validate_prediction_matrix`
- `my_normalize_text`
- `my_prompt_fingerprint`

操作及观察

1. 校验预测矩阵与样本身份，排查训练评测交集.
2. 分别计算分类、生成与 pass-at-k 等指标，记录其分母和适用条件.
3. 比较人工标注与模型裁判，检查分歧、切片及高风险失败；使用既定抽样单位构造区间或配对比较.
4. 汇总成本、时延与质量，核对发布清单及在线实验分配规则.

结果判读及提交

提交逐例结果、分母、置信区间和门禁判定. 构造记录用于检验计分逻辑，不能冒充真实模型评测；统计显著性不等于业务收益.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 7　模型安全

配套文件：`notebook/70_model_safety.ipynb`
教材关联：第 10 章.

实验目标

把风险资产、攻击面、控制和证据逐项关联.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_risk_score`
- `my_sha256`
- `my_evaluate_records`

操作及观察

1. 建立资产与威胁清单，说明风险分级的依据和责任主体.
2. 检查数据、权重及配置完整性，比较正常请求与策略违规样本.
3. 使用授权工具调用入口验证主体、对象和动作边界，分析提示注入与多租户隔离.
4. 检查风险矩阵、发布门禁、日志最小化及事件响应材料.

结果判读及提交

提交允许/拒绝用例、原因及残余风险. 字符串过滤不能代替授权；没有触发已知样本不能证明系统不存在漏洞.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 8　上下文学习

配套文件：`notebook/91_prompt_reasoning.ipynb`
教材关联：第 11 章.

实验目标

控制任务、模型和预算后比较提示策略.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_terms`
- `my_example_score`
- `my_select_examples`

操作及观察

1. 冻结问题集、输出格式及最终答案抽取规则.
2. 对照示例选择、直接回答与推理提示，保存每次提示的完整身份.
3. 在相同预算口径下进行多样本生成与 Self-Consistency 聚合.
4. 比较正确率、格式失败及成本，检查错误抽取和无效样本的分母处理.

结果判读及提交

提交逐题结果、样本数、提示版本和聚合规则. 多数一致不保证事实正确；长推理文本也不构成正确性的独立证明.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 9　检索增强及智能体

配套文件：`notebook/90_llm_applications.ipynb`
教材关联：第 12 章、第 15 章、第 16 章、第 14 章.

实验目标

分离检索证据、模型提案与程序执行.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_tokenize`
- `my_make_document`
- `my_chunk_document`

操作及观察

1. 固定文档和查询，检查切块、召回、上下文预算及来源身份.
2. 比较检索基线与基于证据的生成，逐条核对回答依据和缺证据处理.
3. 检查工具参数与执行白名单，沿固定工作流和 ReAct 路径记录状态变化.
4. 对预算耗尽、无效动作及工具失败设置用例，核对终止条件与保留状态.

结果判读及提交

提交召回证据、工具轨迹和失败归因. 本实验不包含 Codex、OpenClaw 或 Hermes 的安装验收；教材介绍的这些平台不等同于此 Notebook 中的有界执行器.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 10　预训练工程

配套文件：`notebook/40_pre_training.ipynb`
教材关联：第 2 章、第 17 章、第 23 章、第 19 章、第 29 章.

实验目标

验证词元目标、优化步和恢复状态的一致性.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_pack_documents`
- `my_collate`
- `my_build_optimizer`

操作及观察

1. 先检查文档打包边界、标签移位及有效词元计数.
2. 记录 AdamW 参数组和学习率计划，以相同数据顺序建立短程基线.
3. 保存检查点并恢复，比较接续训练与不中断训练的损失、参数及优化器状态.
4. 进入真实语料与 Trainer 路径前冻结数据和模型修订，记录训练预算及固定留出指标.

结果判读及提交

提交恢复对照、步数与词元预算、数据版本. 仅重载权重不等于恢复训练；小规模真实训练也不能据此称为可用基座模型.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 11　后训练方法

配套文件：`notebook/41_post_training.ipynb`
教材关联：第 20 章、第 24 章、第 25 章、第 26 章.

实验目标

分别验证监督区域、偏好概率比和组内优势的数值对象.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_format_prompt`
- `my_encode_sft_record`
- `my_padding_report`

操作及观察

1. 手工标注一条指令回答的有效标签，检查动态 padding 及 packing 后的监督区域.
2. 运行 SFT 基线并记录有效词元归一化方式.
3. 对固定偏好样本计算 DPO 目标，区分策略与参考模型的序列对数概率.
4. 对固定奖励组核对 GRPO 优势、裁剪及零方差处理，再查看 TRL 对象映射.

结果判读及提交

提交监督掩码、偏好计算和组内优势表. 固定奖励数值验证只证明目标计算关系，不等于已经完成在线强化学习训练.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [3, 4]；本实验的运行结果仍须按实际配置单独验证.

实验 12　参数高效微调

配套文件：`notebook/42_peft.ipynb`
教材关联：第 21 章、第 22 章.

实验目标

验证低秩增量的参数范围、合并等价性和行为验收.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_parameter_budget`
- `my_lora_story_color_limit`
- `my_draw_lora_story_cells`

操作及观察

1. 计算低秩参数预算，核对只有预期模块可训练.
2. 训练低秩增量并比较合并前、合并后、拆分及重载的固定输入输出.
3. 进入 Qwen3 身份任务时冻结聊天模板、回答掩码、训练记录与评估探针.
4. 记录适配器制品与底座身份，对照训练前后行为及重载后的固定探针.

结果判读及提交

提交参数清单、数值误差和分切片行为结果. 身份回答改变不等于全部任务能力改善; LoRA 和量化底座训练应分别记录精度与执行条件.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出, 不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [5]; 本实验的运行结果仍须按实际配置单独验证.

框架实践把参数注入、监督目标和适配器制品连接起来, 使用 Transformers 负责模型与分词器、PEFT 负责 LoRA、TRL 负责 SFT 训练 [6, 7, 8]. 配套入口为 `book/workbook/examples/train_adapter.py`, 从仓库根目录运行, 与前面的 Notebook 共享机制目标, 不依赖其内存状态.

准备一个已下载并固定修订的因果聊天模型目录, 以及来源明确的训练与验证 JSONL 文件. 入口要求支持 BF16 的 CUDA 设备; 所需显存取决于基座、长度和激活, 不能仅由 LoRA 参数量估计. 可先使用能够在单卡容纳的小型指令模型验证完整流程, 再扩大规模. 模型目录应包含权重、配置、分词器、聊天模板及相应特殊词元信息.

接口核查基线为 TRL 1.13.0 与 PEFT 0.20.0; 训练依赖与推理引擎分别建环境. 先按目标设备安装兼容的 PyTorch, 再安装并检查训练依赖, 记录解析得到的完整版本集合. 以下操作会安装软件, 应在正式运行实验时执行.

```
python -m venv .venv-train
source .venv-train/bin/activate
python -m pip install "trl==1.13.0" "peft==0.20.0" \
  "accelerate==1.15.0" datasets
python -m pip check
python -m pip freeze > /absolute/run/train-packages.txt
```

`/absolute/run` 须替换为已建立的实验记录目录. 依赖解析成功只说明声明的包约束得到满足, 设备算子、模型加载和反向计算仍需运行验证. 重现实验时复用经验证的锁定环境, 升级则重新核对数据处理与训练行为.

每行包含 `prompt` 与 `completion` 两列, 值为角色消息数组. 下面仅展示记录格式, 训练集应使用经过审校且有使用权限的数据; 验证集按文档、来源或任务划分, 避免同源近重复样本泄漏.

```
{"prompt":[{"role":"user","content":" 请解释梯度累积.
↪  "}],"completion":[{"role":"assistant","content":" 梯度累积在多个微批次上累计
↪  梯度, 再执行一次参数更新. "}]}
```

入口要求最后一条提示来自用户, 且待学习部分恰好是一条助手回答. 它检查训练与验证提示的完全重复、模板后的长度及首批监督目标; 近重复和语义泄漏仍须在数据

准备阶段检查．采用 `completion_only_loss=True`，监督回答部分；`packing=False` 让首次验收保留每条记录的边界．

正式更新前，应从 `trainer.get_train_dataloader()` 取得一个批次，对照 `input_ids` 与 `labels` 解码非-100 位置，确认回答及结束标记受到监督，提示没有被错误计入．源码中的非空检查只能发现没有训练目标的情况，不能替代这项语义检查．EOS 参数应取聊天模板规定的回答结束词元，不应仅凭词元名称猜测．

把路径替换为实际的绝对路径，并从模型模板核实 `EOS_TOKEN`．下面的 100 步用于完整流程验收，不代表已达到某项质量目标；步数、学习率和验证频率应根据数据与验证曲线调整．

```
python book/workbook/examples/train_adapter.py \
  --model /absolute/model-snapshot \
  --train /absolute/data/train.jsonl \
  --eval /absolute/data/validation.jsonl \
  --output /absolute/run/lora-01 \
  --eos-token "$EOS_TOKEN" \
  --steps 100 --max-length 1024 --accumulation 8
```

固定单卡微批次为 1，累积 8 步；LoRA 秩为 16，缩放参数为 32，目标为 PEFT 识别的线性层．先核对打印或保存的可训练参数名称，确认输出头和额外保存层是否符合任务要求．入口拒绝向已有非空输出目录开始新训练；恢复时必须显式指定检查点．

`final/` 保存适配器和分词器，`checkpoint-*` 用于训练恢复，`experiment.json` 记录参数及数据哈希．另行保存基座的不可变修订或文件哈希、驱动与库版本、显存和时间记录．仅有本地路径不能唯一标识基座内容．

在新的进程中加载同一基座，再通过 `PeftModel.from_pretrained` 附加 `final/`．使用相同模板、精度和 `eval()` 模式，比较保存前后固定输入的 logits 或损失，并记录容差；随后在保留测试集上比较微调前后任务质量与能力退化．需要合并时，在可容纳的浮点基座上执行 `merge_and_unload` 并保存新的完整目录，独立验证后再交给推理引擎．量化基座、合并和重新量化的误差应分别测量．

LLaMA-Factory 与 Axolotl 适合把重复的数据处理、训练和导出组织成配置工作流 [9, 10]．迁移时以同一数据与制品为输入，逐项映射下表．字段拼写按所固定的版本查阅，不能只复制不同版本的配置文件．

比较对象	必须保持的语义	验收证据
数据与模板	样本划分、角色、EOS 和回答监督范围	同一样本的词元与标签．
适配参数	目标模块、秩、缩放及额外保存层	可训练参数名称与数量．
优化过程	有效全局批量、更新次数、精度与学习率	配置展开结果与训练日志．
制品	基座身份、适配器与分词器绑定	独立重载和固定探针结果．

TRL 中的 DPO 或 GRPO 训练器对应另一类数据与反馈契约. 偏好训练需要同一提示下的候选偏好, 策略训练还要规定采样和奖励; 把训练器名字替换掉, 不能使 SFT 数据自动成为有效的偏好或强化学习数据.

提交依赖记录、数据划分、监督位置样例、可训练参数、验证损失和独立重载结果. 损失不变时先检查标签、学习率与参数是否获得梯度; 输出异常结束时核对模板和 EOS; 显存不足时分别分析权重、激活、长度与批量. 不得用单条身份回答变化替代保留集评估.

实验 13　训练数据生产

配套文件：`notebook/A20_data_engineering.ipynb`
教材关联：第 18 章、第 23 章、第 20 章.

实验目标

把来源记录追踪到训练分片，同时控制污染和泄漏.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_deep_freeze`
- `my_required_text`
- `my_adapt_document_text`

操作及观察

1. 对不同源 Schema 进行规范化路由，将错误记录隔离并保留原因码.
2. 依次处理规范化、PII、质量判定和去重，比较每一步的保留集合.
3. 核对稳定划分、来源混合与评测污染规则，检查近重复候选的实际相似度.
4. 检查 tokenizer、packing 和 sharding 中的来源区间覆盖，记录输出清单和内容哈希.

结果判读及提交

提交转换血缘、拒绝统计、分片清单与污染检查. 近似去重候选不等于重复事实；同一来源衍生样本不应跨划分泄漏.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 14　推理训练

配套文件：`notebook/A30_reasoning_model.ipynb`
教材关联：第 24 章、第 26 章.

实验目标

区分奖励构造、候选验证和计算分配.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_execute_steps`
- `my_masked_causal_lm_loss`
- `my_verify_candidate`

操作及观察

1. 检查推理轨迹中有效监督位置及最终答案提取.
2. 构造可验证候选，比较步骤执行、拒绝采样和候选筛选结果.
3. 分别计算奖励、优势与裁剪目标，明确验证器能证实的命题范围.
4. 在固定总预算下比较候选数与验证成本，记录停止条件和未解任务.
5. 选择一组可自动核验的题目，固定模型快照、提示、采样参数、输出上限与工具配置，只改变 Notebook 配置中的 `reasoning_effort`. 分别运行 `low`、`medium` 和 `high`；每个档位至少重复三次，并保留随机种子或服务端请求标识.

6. 对每个档位记录最终正确率、验证通过率、可见输出词元、推理词元（若可取得）、端到端延迟、费用、超时、截断与错误. 另做一次输出上限扫描，检查内部推理增加是否挤占可见答案预算.

结果判读及提交

提交验证证据、失败类别、预算账本与推理强度对照表. 若所用模型或运行时不支持某个档位，应保存能力说明或错误响应并将该组标为不适用，不能静默换档. 验证器覆盖有限时，通过验证不能外推为一般正确；候选一致与外部可验证性不同.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 15　资源规划

配套文件：`notebook/A10_resource_planning.ipynb`
教材关联：第 36 章、第 37 章、第 38 章、第 27 章、第 23 章、第 28 章.

实验目标

将显存、计算期限和主机资源分别计量.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_estimate_parameter_breakdown`
- `my_estimate_inference_memory`
- `my_estimate_training_memory`

操作及观察

1. 从参数形状核对参数数目，区分常驻权重与活跃计算.
2. 分别列出推理权重、KV 缓存、工作区以及训练梯度和优化器状态.
3. 改变批次、上下文与精度，观察各项预算如何变化.
4. 将 GPU 数量与期限、CPU、内存、存储及网络需求联合审查.

结果判读及提交

提交分项字节表、单位换算和敏感性分析. 估计值不是实测峰值；未计碎片与通信不能声称配置一定可运行.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 16　训练系统优化

配套文件：`notebook/A40_training_optimization.ipynb`
教材关联：第 36 章、第 19 章、第 21 章、第 22 章、第 28 章、第 29 章.

实验目标

先建立数值正确性，再比较优化收益.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_format_prompt`
- `my_encode_sft_record`
- `my_padding_report`

操作及观察

1. 固定样本、有效词元计数和单步更新基线.
2. 比较梯度累积前后等效批次与参数更新，避免平均方式改变目标.
3. 分别检查混合精度和梯度检查点的数值误差、峰值内存与耗时；记录设备条件.
4. 核对数据并行与张量并行的等价性基线，再分析 ZeRO 和混合并行的通信及内存职责.

结果判读及提交

提交更新误差、同步计时和内存口径. 单进程的两 rank 等价性演算不是多机吞吐实测；首次编译与预热成本应单列.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [11]；本实验的运行结果仍须按实际配置单独验证.

先完成实验 12 的单卡训练链条，再以相同入口研究多设备执行. 本实验选择 DeepSpeed ZeRO-2 作为首次分片运行，随后比较 ZeRO-3 与 FSDP2；Megatron-LM 用于进一步分析模型并行的迁移条件 [12, 13, 14].

准备至少两张支持所选精度的 GPU，记录每卡显存、互联、驱动和通信库. 单机验证通过后再考虑跨节点；跨节点还需核对节点间地址、进程数量、网络和共享制品. LoRA 的优化器状态较小，ZeRO-2 的节省可能有限；研究完整训练状态分片时可使用入口的`--full-finetune`，同时重新核定显存和学习率，不能把两种训练方式混入同一性能对照.

记数据并行设备数为 D_p，每设备微批次为 b，梯度累积次数为 a，则完整批次的样本数为 D_pba. 单卡 $b=1, a=8$ 与两卡 $b=1, a=4$ 均为 8，但尾批次、填充和不同回答长度仍可能改变有效监督词元数. 对照中应保存样本划分、尾批处理及损失归一化方式.

为 DeepSpeed 建立 `/absolute/run/zero2.json`：

```
{
  "bf16": {"enabled": "auto"},
  "zero_optimization": {"stage": 2},
  "gradient_accumulation_steps": "auto",
  "train_micro_batch_size_per_gpu": "auto",
  "train_batch_size": "auto"
}
```

这里的 `auto` 由 Transformers 训练器与 Accelerate 集成解析，实际数值从训练参数取得；这份配置不能作为完全脱离该集成的通用 DeepSpeed 配置 [15]. 优化器和调度器由上层训练配置管理，不在第二处重复定义.

在训练环境中按所用 PyTorch 与加速器安装 DeepSpeed 并保存版本及算子诊断. 启动命令为：

```
torchrun --standalone --nproc_per_node=2 \
  book/workbook/examples/train_adapter.py \
  --model /absolute/model-snapshot \
  --train /absolute/data/train.jsonl \
  --eval /absolute/data/validation.jsonl \
```

```
--output /absolute/run/zero2-01 \
--eos-token "$EOS_TOKEN" --steps 100 \
--max-length 1024 --accumulation 4 \
--deepspeed /absolute/run/zero2.json
```

确认每个进程绑定不同 GPU，打印的世界大小为 2，生效全局批量与设计相符．不要给训练模型额外添加面向推理的 `device_map="auto"`；设备与状态归属由训练运行时管理．

选择一次已完整保存的检查点，以同样的两进程启动命令加入`--resume /absolute/run/zero2-01/checkpoint-40`，并保持原输出目录．该路径须对应实际存在的检查点，不能按示例步数假定保存成功．比较恢复后的步数、数据位置、学习率和下一次更新，核对优化器、调度器及随机状态均按声明恢复．

进一步把 ZeRO 阶段改为 3 时，先记录参数在何时收集及显存峰值，再核查模型保存所需的权重汇总设置．分片训练状态不应直接当作推理权重目录．保存期间参与集合操作的所有进程必须遵循一致的协议；检查点写完后应进行独立加载，部分文件存在不足以证明完整性．

FSDP2 通过 `fully_shard` 与逐参数分片组织执行；其接口机制可参考 PyTorch v2.11.0 的对应源码 [13]，实际安装版本按目标环境验证．采用 Accelerate 集成时，用所安装版本的 `accelerate config --config_file /absolute/run/fsdp.yaml` 生成配置，显式选择 FSDP2、设备数和混合精度，再检查模型模块的包裹边界与检查点策略．旧 FSDP 配置中的字段不保证可直接迁移．

使用 `accelerate launch --config_file /absolute/run/fsdp.yaml` 启动同一训练入口，并去掉 DeepSpeed 配置参数．先用全参数的小规模模型验证分片与恢复，再验证 LoRA 等参数冻结组合．记录所用版本是否支持该模型的冻结方式、梯度检查点和保存路径；不把 FSDP 机制正确等同于每种集成组合都可运行．

Megatron-LM 迁移还要映射模型配置、权重布局、词表、张量与流水线切分，并校验转换前后的固定输入输出．对需要大规模模型并行的任务，它提供另一条工程路径；对已经能用数据并行满足预算的任务，增加转换和并行维度的收益应通过测量证明．

每种策略保留同一训练预算下的验证指标，并报告每设备峰值显存、有效监督词元吞吐、更新耗时、通信等待和恢复耗时．先预热再计量稳定区间，首次算子编译和制品加载单列．通信超时时按进程数、设备绑定、互联和集合调用一致性排查；吞吐下降时区分填充、重计算、通信及存储等待．提交实际运行的策略组合和原始日志，资源不足的迁移项按实际状态记录．

实验 17　推理服务

配套文件： `notebook/60_inference_deployment.ipynb`
教材关联： 第 30 章、第 31 章、第 32 章.

实验目标

用明确的对象身份和状态转换分析服务可靠性.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_plan_replicas`

操作及观察

1. 检查不可变制品、配置及请求 Schema，记录模型修订身份.
2. 沿 Gateway、Router、Scheduler、Worker 逐层追踪请求与资源责任.
3. 用给定负载计算队列和副本需求，逐项列出容量假设.
4. 检查取消、背压、故障与回滚路径，区分已接收、已执行和结果未知.

结果判读及提交

提交制品清单、请求时序、容量推导和恢复判据. 配置验证与容量估算不是线上压测，也不能证明服务已经部署.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

先固定模型权重、分词器、聊天模板、精度与最大输入输出长度. 框架实践以 vLLM 完成服务链条，再把同一制品迁移到 SGLang；TensorRT-LLM 作为平台与制品约束的对照 [16, 17, 18]. 推理引擎接收完整模型目录；若实验 12 生成的是适配器，应先完成合并和独立验证，或另外核查引擎的动态适配器支持.

命令接口分别按 vLLM 0.29.0 与 SGLang 0.5.19 核对. 为两个引擎使用独立环境，按各自发行版的加速器安装说明准备软件；不能假定它们与训练环境共享同一 PyTorch、CUDA 和算子依赖. 记录包锁或镜像摘要、驱动、GPU 型号和引擎启动参数. 模型与数据由实验执行者预先准备，书籍编译不触发下载.

先用单卡可容纳的模型验证请求路径，之后再研究量化与多卡并行. 可用显存不仅要容纳权重，还要留出缓存、激活与工作空间；标称上下文长度也不等于既定并发下可以承载的长度.

在 vLLM 环境中启动本地服务，保留启动日志：

```
vllm serve /absolute/model-snapshot \
  --host 127.0.0.1 --port 8000 \
  --generation-config vllm
```

`--generation-config vllm` 使对照不隐式继承模型目录中的生成默认值；请求仍需显式给出采样与停止条件 [16]. 先读取`/v1/models` 确认实际服务模型名称，再向`/v1/chat/completions` 发送请求. 以下客户端只使用 Python 标准库；将其保存为实验记录中的 `request.py`，首次以端口 8000 运行.

```
import json
import sys
import urllib.request

base = sys.argv[1].rstrip("/")
opener = urllib.request.build_opener(
    urllib.request.ProxyHandler({}))
with opener.open(base + "/v1/models", timeout=30) as r:
    models = json.load(r)["data"]
if len(models) != 1:
    raise ValueError("Select and bind one served model explicitly")
body = {
    "model": models[0]["id"],
    "messages": [{"role": "user", "content": " 请简述梯度累积. "}],
    "temperature": 0,
    "max_tokens": 128,
    "stream": False,
}
request = urllib.request.Request(
    base + "/v1/chat/completions",
```

```
    data=json.dumps(body, ensure_ascii=False).encode("utf-8"),
    headers={"Content-Type": "application/json"})
with opener.open(request, timeout=120) as r:
    result = json.load(r)
assert result.get("choices"), result
print(json.dumps(result, ensure_ascii=False, indent=2))


python /absolute/run/request.py http://127.0.0.1:8000
```

保存完整响应，检查模型身份、内容、停止原因和词元计数．状态码成功只能证明请求被处理，回答是否正确仍需任务判据．上述非流式客户端不能测量首词元或逐词元延迟；这些指标需要流式事件时间或服务器遥测．对外服务时，将认证、TLS、限流及访问控制接到经过验证的入口，不能直接把本地检查地址替换为公网监听便称为部署完成．

推理模型还要核查“模型支持”和“引擎接口支持”两个条件．先从固定版本的模型卡与引擎请求文档确认推理强度的字段名、允许值、默认值和用量字段，再向测试服务发送一个支持值和一个故意无效的值．保存原始请求、响应和服务日志；只有错误可识别、响应能回显或日志能证明最终生效值时，才把该参数纳入跨引擎对照．OpenAI 兼容路径不保证实现 OpenAI 全部字段，不能看到`/v1/chat/completions`便假定`reasoning_effort`有效．

停止前一引擎并确认显存释放，再在 SGLang 环境启动相同模型目录：

```
python -m sglang.launch_server \
  --model-path /absolute/model-snapshot \
  --host 127.0.0.1 --port 30000
python /absolute/run/request.py http://127.0.0.1:30000
```

参数定义与启动路径见相应版本源码和官方请求说明 [19, 20]．客户端读取实际模型名称，避免把端口切换误当作全部迁移工作．比较模板渲染后的输入词元、特殊词元、EOS 及有效生成参数；相同消息 JSON 可能在不同配置下形成不同输入．

对确定性设置仍应声明允许的数值或答案差异．内核、批组成和浮点规约可能改变生成轨迹；不能仅凭`temperature=0`要求不同引擎逐字相同．若引擎支持所需 logprobs，可先比较固定前缀的分数，再用保留测试集评估任务质量．

构造来自实际目标负载的短、中、长输入组，分别准备低共享前缀和高共享前缀两组，固定期望输出长度与停止规则．对于支持推理强度的模型，再按实际支持档位分层；每轮只使用一个档位，并记录请求值与最终生效值．先预热，再逐级增加并发或到达率；使用两者共同支持的接口回放同一请求清单，保存原始事件和错误．只改变一个因素，避免把量化、并行数、推理强度和请求分布同时更换．

指标	采集位置	判读要求
首词元与逐词元延迟	流式客户端或引擎遥测	区分排队、预填充与解码；网络事件未必一词元一事件.
端到端延迟与尾延迟	请求发送至完成	报告分位数、样本量和超时.
有效吞吐	固定测量窗口	统计满足质量与延迟门槛的请求或输出词元.
缓存与显存	引擎与设备指标	区分模型、缓存与工作空间，注明冷热条件.
失败比例	客户端与服务日志	保留超时、取消、资源不足和异常停止.
推理用量	响应用量与引擎遥测	分开记录内部推理与可见输出；缺少字段时注明不可观测.

只看成功请求的平均延迟会隐藏丢弃与超时. 闭环固定并发和开环固定到达率也有不同的排队含义，应分别说明. 比较前缀缓存时，冷启动成本与稳定命中收益分开计量，并核对缓存开启状态和模型身份.

在可控实验服务上取消长请求，观察连接结束后缓存是否回收；发出超出长度限制的请求，核查错误是否可解释；终止一个工作进程，记录检测和恢复时间. 故障之后再次验证模型身份、固定探针和容量. 对 TensorRT-LLM 的迁移，另行核对 GPU 支持、执行后端、模型转换或加载步骤、精度与构建成本，不能把某个后端的要求推广到整个项目.

提交制品哈希、引擎与硬件版本、完整启动配置、固定请求清单、能力探测结果、质量对照、分档负载曲线和故障结果. 首次请求、稳定运行与恢复分别标记. 框架声明的特性可以作为选型条件，实验结论必须来自本次保存的运行证据.

实验 18　推理优化

配套文件：`notebook/A50_inference_optimization.ipynb`
教材关联：第 37 章、第 30 章、第 33 章、第 34 章、第 28 章、第 31 章、第 32 章.

实验目标

把数值等价性、内存节省和时延收益分别验证.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_format_prompt`
- `my_encode_sft_record`
- `my_move_batch`

操作及观察

1. 固定输入与模型，比较朴素生成和 KV 缓存生成的词元及输出差异.
2. 跟踪缓存长度、头数和内存估算，检查左 padding、分桶及批处理状态.
3. 比较 SDPA、量化和编译路径时记录设备、预热、同步、输入形状及版本.
4. 分别报告 TTFT、TPOT 和吞吐，审查连续批处理、推测解码及并行部署的适用前提.

结果判读及提交

提交缓存轨迹、等价误差与分阶段时间. 状态转换演算不代表真实服务压测；前缀缓存还需考虑会话身份和版本隔离.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [21, 22]；本实验的运行结果仍须按实际配置单独验证.

实验 19　模型压缩

配套文件：`notebook/A60_model_compression.ipynb`
教材关联：第 34 章.

实验目标

区分参数表示变化和执行路径变化.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_global_magnitude_prune`
- `my_prune_mlp_hidden`
- `my_factorize_linear_svd`

操作及观察

1. 固定基线权重与评价输入，分别应用非结构化剪枝和结构化剪枝.
2. 比较低秩近似的矩阵形状与误差，记录秩和参数预算.
3. 核对量化尺度、量化值与反量化误差；检查蒸馏目标的教师/学生对象.
4. 组合方法时逐步增加变量，记录质量、内存、制品和实际执行后端.

结果判读及提交

提交逐阶段差异及制品配置. 零值权重并不自动减少稠密算子耗时；模拟量化误差不能代替低比特内核速度测试.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 20　掩码语言建模

配套文件： `notebook/E30_nlp_bert.ipynb`
教材关联： 第 4 章、第 35 章.

实验目标

核对词元、嵌入、注意力与预训练目标之间的契约.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_wordpiece`
- `my_encode`

操作及观察

1. 检查 WordPiece 分词结果与特殊词元，记录三类嵌入的组合.
2. 逐层核对双向注意力形状，与因果掩码对照可见位置.
3. 检查 MLM 被选位置及标签，区分预训练目标和下游分类头.
4. 迁移到 Transformers 时比较配置、输入字段与输出对象.

结果判读及提交

提交可见性对照和目标位置表. 分类准确率不能直接验证 MLM 实现；双向编码器不能简单按自回归方式生成下一词元.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [23]；本实验的运行结果仍须按实际配置单独验证.

实验 21　开放模型制品审计

配套文件：`notebook/E10_open_model.ipynb`
教材关联：第 36 章、第 38 章.

实验目标

以文件证据核对结构，而非按品牌推断能力.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_rms_norm`
- `my_top_k_route`
- `my_kv_cache_elements`

操作及观察

1. 固定仓库修订及本地根目录，列出配置、分词器、权重索引和相关代码.
2. 比较配置中的维度、注意力头与专家字段，核对权重头部和张量结构.
3. 检查特殊词元、聊天模板、适配器和量化元数据的一致性.
4. 导出审计清单，明确静态可证事实、缺失证据及需运行才能确认的项目.

结果判读及提交

提交审计清单与文件哈希. 仅解析配置不证明前向计算通过；不应为了查看仓库而默认执行自定义远程代码.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 22　视觉 Transformer

配套文件： `notebook/E40_cv_vit.ipynb`
教材关联： 第 39 章.

实验目标

验证 patch 排列和视觉编码器接口.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_patchify`

操作及观察

1. 使用可追踪的图像数组检查 patchify 的像素顺序.
2. 比较补丁切分加线性投影与对应卷积投影的结果.
3. 核对 CLS、位置编码及编码器输出形状，检查输入分辨率与补丁数.
4. 对照 torchvision 或 Transformers 接口，迁移分类头时记录预处理和权重来源.

结果判读及提交

提交补丁可视化、等价误差和预处理配置. 形状一致不能排除像素顺序错误；改变分辨率需重新处理位置表示.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [24]；本实验的运行结果仍须按实际配置单独验证.

实验 23　多模态数据

配套文件： `notebook/E20_multimodal_llm.ipynb`
教材关联： 第 40 章.

实验目标

验证视觉表示如何进入语言模型的序列与监督区域.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_draw_connector_block`
- `my_draw_connector_arrow`

操作及观察

1. 记录图像输入与 patch token 的形状，区分视觉分辨率与词元长度.
2. 对照 Projector 和 Resampler 的维度与长度变换.
3. 组装视觉和文本序列，逐项检查占位符、注意力可见性及标签掩码.
4. 比较冻结与解冻阶段的可训练模块，审查迁移到标准处理器时的协议变化.

结果判读及提交

提交形状流、掩码与参数范围. 张量连接正确不证明视觉问答能力；模型图谱快照不能当作持续更新的排名.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

实验 24　条件视觉生成

配套文件： `notebook/E50_cv_diffusion.ipynb`
教材关联： 第 41 章、第 42 章.

实验目标

核对前向噪声过程、噪声预测和反向采样的参数化.

准备及定位

先阅读本册实验总则，确认 Notebook 中的依赖与资产单元. 原理计算、库接口迁移和真实模型训练可能具有不同资源要求，应分别记录设备、版本及运行范围.

源码定位可从下列函数开始；应结合前置数据与配置单元阅读.

- `my_linear_beta_schedule`
- `my_extract`
- `my_add_noise`

操作及观察

1. 固定随机种子，比较不同时间步的加噪结果和噪声计划.
2. 核对时间嵌入、噪声预测器的输入输出及训练目标.
3. 比较有条件与无条件预测，记录 CFG 系数和 VAE 潜变量尺度.
4. 检查反向均值与采样器接口，迁移 Diffusers 时核对预测类型、时间步和调度器.

结果判读及提交

提交噪声轨迹、损失定义及采样配置. 单次生成观感不构成分布质量评测；不同预测参数化不能直接共用同一反向公式.

提交时附上所执行单元范围、环境与制品身份、原始结果及失败说明. 将未运行项目明确列出，不能用 Notebook 中已有输出替代本次执行证据.

相关方法的原始定义与适用前提见 [25]；本实验的运行结果仍须按实际配置单独验证.

第三部分

参考解析

本部分按教材章号与题号汇集习题的参考解析. 计算与推导题列出必要前提、关键步骤和结果；设计与讨论题给出一种有依据的分析路径，其他满足题设与约束的方案也可能成立.

扩展题的解析在外侧页边沿用原题的危险弯道标记，表示首次阅读可暂时跳过.

第一章　绪论

1.1　条件分布仍须归一化，故 $p(v \mid b) = 0.3$. 四条路径为 $0.6 \times 0.5 = 0.30$、0.30、$0.4 \times 0.7 = 0.28$、$0.4 \times 0.3 = 0.12$. 最大路径改为经过 a 的两条并列路径，不再经过 b. 仅改变一个条件概率而不相应调整另一项，会破坏本题的概率模型.

1.2　设同一文本编码长度为 N_i，每秒生成 r_i 个词元，纯解码时间约为 $\dfrac{N_i}{r_i}$. 较大的 r_i 可能伴随更大的 N_i；还应计入分词、预填充和调度. 应固定原始文本、生成质量与停止条件，比较完整请求时延、字符或字节处理量，并记录两套分词器和输入输出长度.

1.3　依次检查正确片段是否进入实际上下文、是否被截断或错误序列化、来源版本与时间是否匹配、答案是否忠实使用证据，以及引用是否对应具体论断. 针对故障可调整切分、重排、上下文编排、证据校验或拒答条件. 扩大参数量不能修复丢失片段、错引用或权限过滤错误；修复后用同一问题及干扰变体回归.

1.4　同时记录原始词元数、有效监督数、标签掩码、样本来源权重、批量及更新次数. 若缩短序列或增加被屏蔽位置，原始词元每秒提高不代表有效学习加速. 应保持目标位置和权重一致，比较有效监督词元每秒、单位更新损失及独立验证质量；若目标改变，明确作为新训练方案比较.

第二章　数学基础

2.1　令正确率为 0.8，则通过概率为 $0.8 \times 0.9 + 0.2 \times 0.1 = 0.74$，后验正确率为 $\dfrac{0.72}{0.74} = \dfrac{36}{37} \approx 0.972973$. 验证器条件接受率未变，但错误候选的基数降低，因此通过集合更纯；不能只凭敏感度 0.9 确定后验.

2.2　交叉熵为 $H(p, q) = -\log(\dfrac{1}{2}) = \log 2$，所以有限. 反向散度中，$q_2 > 0$ 而 $p_2 = 0$，第二项为

$$\frac{1}{2} \log \frac{\frac{1}{2}}{0} = +\infty.$$

因此 $D_{\mathrm{KL}}(q \| p) = +\infty$. 正向散度则为 $D_{\mathrm{KL}}(p \| q) = \log 2$，因为 $0 \log(\dfrac{0}{q_2})$ 按连续延拓取零.

2.3

$$H(p) = -.6\log.6 - .4\log.4 \approx .673012, \quad H(p,q) = -.6\log.7 - .4\log.3 \approx .695594.$$

两者之差为 $.6\log(\frac{6}{7}) + .4\log(\frac{4}{3}) \approx .0225824 = D_{\mathrm{KL}}(p\|q)$. 反向为 $.7\log(\frac{7}{6}) + .3\log(\frac{3}{4}) \approx .0216009$；一般不相等.

2.4 设原四项损失为 $\ell_1, \dots, \ell_4$. 若目标、条件上下文、位置协议、参数、随机状态和权重均不变，切片后两段的损失和仍为四项之和；按有效词元合并平均仍为其四分之一. 若第二段丢失前两词元上下文，$p(x_3 \mid x_1, x_2)$ 变为另一条件分布，原有等式无须成立. 还应避免额外 EOS 和重复监督改变目标集合.

2.5 $\dfrac{e^{z_j+1000}}{\sum_k e^{z_k+1000}} = \dfrac{e^{z_j}}{\sum_k e^{z_k}}$，理论概率不变. 直接指数可能溢出，产生无穷比无穷. 稳定计算先减去行最大值；对数损失使用 log-sum-exp 而非先求可能舍入为零的概率再取对数.

2.6 对 q 求导得

$$L'(q) = -\frac{wp}{q} + \frac{1-p}{1-q}, \qquad L''(q) = \frac{wp}{q^2} + \frac{1-p}{(1-q)^2} > 0.$$

令一阶导数为零，得到唯一最优解及其反变换

$$q^* = \frac{wp}{wp + 1 - p}, \qquad p = \frac{q^*}{w - (w-1)q^*}.$$

代入题设有 $q^* = 0.5$，而真实正类概率为 0.1. 加权改变了损失最优输出的含义；固定正权重时该变换对 p 单调，但数值校准与排序是不同性质. 反变换依赖权重已知、条件风险达到理想最优以及数据分布匹配. 有限容量、优化误差或分布变化下，仍应在独立数据上检查校准，并按实际误判代价选择阈值.

第三章　神经网络基础

3.1 $Z_{ij} = \sum_{k=1}^{4} X_{ik}W_{kj} + b_j$，$G \in \mathbb{R}^{3\times 2}$. 由 $\mathrm{d}Z = (\mathrm{d}X)W + X\,\mathrm{d}W + \mathbf{1}_3\,\mathrm{d}b^{\mathsf{T}}$ 得 $\overline{X} = GW^{\mathsf{T}} \in \mathbb{R}^{3\times 4}$、$\overline{W} = X^{\mathsf{T}}G \in \mathbb{R}^{4\times 2}$、$\overline{b} = G^{\mathsf{T}}\mathbf{1}_3 \in \mathbb{R}^2$. 同一偏置在三行复用，其梯度必须对三条路径求和.

3.2 记 $R = XW - T$. $\mathrm{d}\mathcal{L} = \mathrm{tr}[R^{\mathsf{T}}((\mathrm{d}X)W + X\,\mathrm{d}W)]$，循环移动迹中因子可得 $\nabla_W\mathcal{L} = X^{\mathsf{T}}R$、$\nabla_X\mathcal{L} = RW^{\mathsf{T}}$. 若对 B 个样本平均，目标乘 $\frac{1}{B}$，两个梯度也乘 $\frac{1}{B}$；不能额外再除输出宽度，除非所定义目标还按元素平均.

3.3

$$J = \operatorname{diag}(q) - qq^{\mathsf{T}} = \begin{bmatrix} \frac{1}{4} & -\frac{1}{6} & -\frac{1}{12} \\ -\frac{1}{6} & \frac{2}{9} & -\frac{1}{18} \\ -\frac{1}{12} & -\frac{1}{18} & \frac{5}{36} \end{bmatrix}.$$

每行和为零. 任选 $v = (1,2,3)^{\mathsf{T}}$, $q^{\mathsf{T}}v = \frac{5}{3}$, 故 $Jv = q \odot (v - \frac{5}{3}\mathbf{1}) = (-\frac{1}{3}, \frac{1}{9}, \frac{2}{9})^{\mathsf{T}}$, 与逐行矩阵乘法相同. 题目未指定向量, 这是一种可复核选择.

3.4 设 $s = \sum_j y_j$, 展开损失为 $-\sum_j y_j z_j + s \log \sum_k e^{z_k}$, 故梯度为 $sq - y$. 本题 $s = 2$, 得到 $2q - (1,1,0)$. Softmax 类别共享一个分母; 三个独立二元任务使用各自 Sigmoid, 梯度是 $\sigma(z_j) - y_j$, 不存在该互斥归一化约束.

3.5 $\bar{b}_{1j1} = \sum_{i=1}^{2} \sum_{k=1}^{4} \overline{Z}_{ijk}$, 即对批量轴及最后一轴求和, 保留维度. 若 $\mathcal{L} = \operatorname{mean}(Z)$, 则每个 $\overline{Z}_{ijk} = \frac{1}{24}$, 偏置梯度为 $\frac{8}{24} = \frac{1}{3}$. 若平均的是后续逐元素损失, 其局部导数先乘 $\frac{1}{24}$ 再沿广播轴求和.

3.6 第 1 行梯度为 $(1,0) + (2,3) = (3,3)$, 第 3 行为 $(0,1) + (-1,2) = (-1,3)$, 未出现行均为零. 覆盖写入只保留最后一项, 将得到错误的 $(2,3)$ 和 $(-1,2)$. 反向应执行按编号的散射求和, 而非赋值覆盖.

3.7 将 $W_{1,12}$ 增加 t, 两个相关预激活变为 $1+t$ 与 $3-2t$. 例如 $|t| < \frac{1}{2}$ 保证激活分支不变. 正确类别分数差分别为 $2-t$、$3-2t$, 故

$$f(t) = \tfrac{1}{2}\log(1 + e^{-2+t}) + \tfrac{1}{2}\log(1 + e^{-3+2t}), \quad f'(0) = \frac{a}{2} + b \approx 0.1070273.$$

求偏导时其他参数固定; 同时更新其他参数将不再是这个单变量函数.

3.8 中心差分为 $\frac{[f(\varepsilon) - f(-\varepsilon)]}{2\varepsilon} = \frac{1}{2}$. 零点不存在普通导数, 差分跨越左右两分支, 所以不能据此否定框架取零的次梯度约定. 在 $x = 1$ 附近取 $0 < \varepsilon < 1$ 可保持两点都在正半轴, 差分精确为 1; 浮点计算还需避免扰动过小产生消减误差.

3.9 记有效损失和为 S, 填充后位置数 N_p、有效数 $N_v > 0$, 则 $L_p = \frac{S}{N_p}$、$L_v = \frac{S}{N_v}$, 并有 $\nabla L_p = (\frac{N_v}{N_p})\nabla L_v$. 填充比例变动会改变前者梯度尺度; 若不同批次有效长度不同, 也改变批次贡献. 两种函数都可通过各自有限差分测试, 该测试只检验导数实现, 不证明目标相同.

3.10 差分一致支持所实现标量函数的局部导数计算正确, 不能证明该函数符合预测任务. 对序列 (a,b,c), 下一词元预测使用输入 (a,b)、标签 (b,c); 若标签为 (a,b), 当

前词元已出现在该位置输入中，模型可能学到复制. 应手算一个短序列的目标位置与有效词元掩码，逐位置核对损失，并通过改变未来词元检查因果隔离. 语义检查通过后，再在光滑点比较中间量与梯度. ReLU 折点处经典导数不存在，跨越两侧的中心差分可与框架选用的次梯度不同；应换用远离折点的输入，不能据此单独判定反向传播实现错误.

第四章　文本表示

4.1　`aaaa` 中 aa 相邻出现 3 次，权重二贡献 6；`aab` 再贡献 1 次 aa 和 1 次 ab，故 aa 计 7、ab 计 1. 令 aa 合并为 A，左到右非重叠替换得到 `AA`（权重二）和 `Ab`. 总词元数从 11 降到 6，减少 5 而非 7，因为同一轮重叠的 aa 不能同时替换.

4.2　按 ab、ac、abab 三条规则，边界依次为 (a,b,a,c,a,b,a,b)、(ab,a,c,ab,ab)、(ab,ac,ab,ab)、$(ab,ac,abab)$. 交换前两条时先得到 (a,b,ac,a,b,a,b)，再得到同一 (ab,ac,ab,ab)，最终不变. 这两种基础对在该输入不重叠；一般反例为 `abc` 上 ab 与 bc 两规则，先后分别得到 (ab,c) 和 (a,bc). 单个输入相同不足以证明任意规则可交换.

4.3　选词表含 `ab`,`a`,`##bc`，不含 `##c`，输入为 `abc`. 贪心先取最长词首 ab，剩 c 无续接项而失败；若回溯改取 a，则可接 `##bc`. 标准贪心不会自动执行该回溯；加入回溯改变了编码算法. 具体 UNK 返回粒度还需按所用实现约定.

4.4　对 `abab`，路径为 (a,b,a,b)、(ab,a,b)、(a,b,ab)、(a,ba,b)、(ab,ab)，权重依次为 $.0016,.016,.016,.008,.16$，总权重 .2016，最大权重 .16. 前向值为 $\alpha_0=1,\alpha_1=.2,\alpha_2=.44,\alpha_3=.128,\alpha_4=.2016$，最后一步 $.128\times.2+.44\times.4=.2016$. 总和等于最大值要求其余路径权重全为零，或仅有一条正权重路径.

4.5　更新参数为 $q(a),q(b),q(ab),q(ba)=\dfrac{17,1,10,5}{33}$. 三路径权重在共同分母 33^3 下为 5610、2805、289，总和 8704；后验因此为 $\dfrac{5610}{8704}$、$\dfrac{2805}{8704}$、$\dfrac{289}{8704}$，约为 .64453, .32227, .03320. 一条样本只约束该字符串的覆盖，无法识别其他语言的字符、边界与频率，必须使用具有来源覆盖和独立验证的语料.

4.6　若 $N(x)=N(y)$ 且 $x\neq y$，任何只接收 $N(x)$ 的确定解码器 D 都只能给同一结果，因而不可能同时满足 $D(N(x))=x$ 和 $D(N(y))=y$. 代码中的大小写标识 `A`/`a`、数学中普通负号与特定运算符或大小写变量可能具有不同含义. 即使某种合并适用于检索，也不能在需原文恢复的流程中无记录地应用.

4.7　固定宽度和批量，密集注意力主要项按 T^2 变化，为原来的 $.8^2=.64$；全词表投影按 TV 变化，为 $.8\times2=1.6$. 总时延还受 FFN/注意力投影、缓存读写、矩阵内核效率、词表并行通信、批量与序列调度影响；参数显存也因词表扩大而增长.

4.8 取显式移位后的七位置输入 x、目标 y，以及输入有效性掩码 m_{in} 和仅回答损失掩码 m_{loss}：

$$\begin{aligned} x &= (\text{BOS}, p_1, p_2, a_1, a_2, \text{EOS}, \text{PAD}), \\ y &= (p_1, p_2, a_1, a_2, \text{EOS}, \text{PAD}, \text{PAD}), \\ m_{\text{in}} &= (1, 1, 1, 1, 1, 1, 0), \\ m_{\text{loss}} &= (0, 0, 1, 1, 1, 0, 0). \end{aligned}$$

第 5 个目标是真实 EOS，应计损失；最后两个是填充，应忽略. 即使编号相同也必须根据来源位置区分，不能按编号批量删除 EOS 监督.

4.9 先比较原文规范化与预分词，再比合并规则、词表字节串到 ID 映射、特殊词元和模板，最后比 padding、offset 和标签策略. ID 变化会错选嵌入行、输出头类别和适配器语义，长度变化影响缓存及预算. 若仅作词表置换，必须同步置换输入输出参数及相关资产；一般重新分词不能用简单行置换恢复，需重新训练或适配并独立评估.

4.10 输入编号 $(3, 1, 3)$ 的独热表示与查表结果为

$$S = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad SE = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

仅交换 E 的第 1、3 行得到 E'，则

$$E' = \begin{bmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix}, \qquad SE' = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

三个输出位置均发生改变. 若同时把旧编号 1 映射到新编号 3、旧编号 3 映射到新编号 1，查表恢复原结果；输出词表及标签也须同步保持同一映射.

4.11 若输入为 $X = \alpha SE$、输出分数 $Z = HE^{\mathsf{T}}$，则 $\overline{E} = \alpha S^{\mathsf{T}}\overline{X} + \overline{Z}^{\mathsf{T}}H$. 取 $\alpha = 2$ 只在第一条链的局部导数显式出现 2；输出路径局部映射没有该乘数. H 本身依赖 α 及整个主干，因此数值上两条梯度都可能改变；不能冻结 H 的推导与完整训练中的依赖混淆.

4.12 原向量内积 $1 - 1 = 0$，余弦为 0. 变换后为 $(10, 1), (10, -1)$，内积 99，两者范数均为 $\sqrt{101}$，余弦为 $\dfrac{99}{101} \approx .980198$. 可逆性只保证信息可恢复；内积需要正交变换才普遍保持，余弦还允许共同的非零等比例尺度.

第五章　注意力机制

5.1　$QK^\top$ 为 5×7，逐键归一化 A 仍为 5×7，AV 为 5×3. 每个查询需要一组对七个键的概率；沿查询轴归一化使每个键在五个查询间分配质量，通常不能保证某个查询权重和为 1，定义了另一运算.

5.2　每个独立乘积 $q_r k_r$ 均值 0、方差 $4\times 4=16$，跨维协方差为零，故点积方差 $16d_k$. 除 $\sqrt{d_k}$ 后方差除以 d_k，得到 16. 缩放消除了维数因子，不自动把输入尺度标准化.

5.3　第一行分数为 $(c,0,c)$，$c=\dfrac{1}{\sqrt{2}}$. 令 $r=\dfrac{e^c}{2e^c+1}$、$s=\dfrac{1}{2e^c+1}$，权重 $(r,s,r)\approx(.401112,.197776,.401112)$，输出为 $(2r,2s+2r)\approx(.802224,1.197776)$. 第三个值置零但 QK 固定时，输出变为 $(r,2s)\approx(.401112,.395552)$；权重不变，差值是原来的 rV_3.

5.4　$J=\operatorname{diag}(a)-aa^\top$，所以 $J\mathbf{1}=a-a(a^\top\mathbf{1})=0$. 对分数加常数 c 时指数分子分母共同乘 e^c，概率精确不变；全一方向是该不变性的微分表示. 雅可比因此至少有一个零特征方向.

5.5　令 $\alpha=\dfrac{1}{1+e^c}$. 第二查询为 $(0,1)$，分数梯度为 $(\alpha(1-\alpha),-\alpha(1-\alpha))$，故 $\overline{K}_1=(0,\dfrac{\alpha(1-\alpha)}{\sqrt{2}})$、$\overline{K}_2=-\overline{K}_1$. 仅扰动 $K_{1,2}$ 为 t，固定 Q、V 及其余 K，损失变为 $f(t)=\dfrac{1}{[1+\exp(c-\frac{t}{\sqrt{2}})]}$. 中心差分 $\dfrac{[f(\varepsilon)-f(-\varepsilon)]}{2\varepsilon}$ 趋向 .1563986；这不是同时改变自注意力源 X 的总导数.

5.6　唯一查询处于绝对位置 8，允许集合为 $\{0,1,\dots,8\}$. 左上对齐的 1×9 下三角仅保留键 0，错误删除 1 至 8. 应按绝对位置比较，或使用已验证具有正确右下对齐语义的接口，不能仅根据矩阵形状生成掩码.

5.7　每层需 $2\times 2048\times 6\times 64\times 2=3145728$ 字节，即 3MiB. MHA 把 KV 头改为 24 得到 12MiB；GQA 为四分之一. 未计分页、尺度、对齐及其他状态，也不意味总推理时延必然缩短四倍.

5.8　相同 A 不同 V：一个键权重 1 时分别取 $V_1=0,V_2=1$，输出不同. 相同 K、V 不同 Q：取标量键 $(0,1)$、值 $(0,1)$，查询分别 0 和 1，输出分别 $\dfrac{1}{2}$ 和 $\dfrac{e}{1+e}$. 反之，特定退化 V 也可能使不同权重输出相同，所以“不同”也非必然.

5.9　条件于固定 A，$S=\displaystyle\sum_j \frac{m_j A_{ij}}{q}$，其中独立 $m_j\sim\text{Bernoulli}(q)$. $\mathbb{E}S=\displaystyle\sum_j A_{ij}=1$，方差为 $\displaystyle\sum_j \frac{A_{ij}^2 q(1-q)}{q^2}=(1-q)\sum_j\frac{A_{ij}^2}{q}$. 单次掩码下 S 不必为 1；诊断时应区分 Softmax 之前、之后及 Dropout 之后的张量.

5.10　一个需回答但检索候选全被权限过滤的有效查询，其允许集为空，应报无可用键或按显式协议回退；不能伪造注意力分布. 一个只为矩形批次补齐的无效查询可直

接输出零并屏蔽损失. 把全行设成同一有限负数会归一化为均匀权重；全为负无穷直接求 Softmax 又可能产生 NaN，两者都不能替代语义判断.

5.11 键值投影的输出宽度和 KV 缓存元素数均变为原来的四分之一；每个查询头仍对历史位置计算注意力分数，因此固定序列长度时分数数量不变. 查询投影、输出投影与前馈计算也不会随 KV 头数按同比例缩小. 实际访存收益依赖内核是否复用共享 KV，端到端时间还包含调度与其他算子. 比较时应固定硬件、精度、输入输出长度分布、并发和质量要求，分别测量缓存峰值、预填充及解码时间. 结构变化通常需要相应权重适配和质量评估，不能把现有 MHA 权重视作无损可替换的 GQA 权重.

第六章　Transformer 网络结构

6.1 没有激活时 $(xW_1+b_1)W_2+b_2=x(W_1W_2)+(b_1W_2+b_2)$. 在固定 ReLU 激活区域令 $D=\operatorname{diag}(\mathbf{1}[xW_1+b_1>0])$，则行向量扰动满足 $\mathrm{d}F=\mathrm{d}x\,W_1DW_2$，列雅可比为 $(W_1DW_2)^{\mathsf{T}}$. 跨越折点会改变 D，因此不同区域具有不同仿射参数.

6.2 采用精确 $\mathrm{GELU}(u)=u\Phi(u)$，其导数为 $\Phi(u)+u\phi(u)$；SiLU 为 $u\sigma(u)$，导数为 $\sigma(u)+u\sigma(u)(1-\sigma(u))$. 在 $u=-1,0,1$，GELU 值约 $(-.158655,0,.841345)$，导数 $(-.083315,.5,1.083315)$；SiLU 值约 $(-.268941,0,.731059)$，导数 $(.072329,.5,.927671)$. 负区间允许负输出，且两者局部斜率不同. 光滑性不能独自决定训练质量，还取决于尺度、结构、数据和优化.

6.3 普通无偏置 FFN 参数 $2df$，门控 FFN 参数 $3dg$，故 $g=\dfrac{2f}{3}=\dfrac{8192}{3}\approx 2730.667$. 256 倍数向下为 2560、向上 2816；相对原 $8192d$ 参数分别少 $512d$ (6.25%) 和多 $256d$ (3.125%). 计算效率还受内核对齐与实际矩阵形状影响，不能仅按参数差判断快慢.

6.4 令 $P=I-\dfrac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{d}$、$c=Px$、$s=\sqrt{\dfrac{c^{\mathsf{T}}c}{d}+\epsilon}$. 有 $\mathrm{d}c=P\,\mathrm{d}x$、$\mathrm{d}s=\dfrac{c^{\mathsf{T}}\,\mathrm{d}x}{ds}$，故无仿射输出雅可比为 $J=\dfrac{P}{s}-\dfrac{cc^{\mathsf{T}}}{ds^3}$；有增益时左乘 $\operatorname{diag}(\gamma)$. 因 $P\mathbf{1}=0$ 且 $c^{\mathsf{T}}\mathbf{1}=0$，$J\mathbf{1}=0$，对应整体平移不改变中心化表示.

6.5 令 $x=(1,2,3)$，归一化均不含可训练仿射参数. 其结果为

$$\mathrm{LN}(x)=\frac{(-1,0,1)}{\sqrt{\frac{2}{3}+\epsilon}},$$

$$\mathrm{RMSNorm}(x)=\frac{(1,2,3)}{\sqrt{\frac{14}{3}+\epsilon}}.$$

各坐标整体加十后，令 $x'=(11,12,13)$. LN 不变，RMSNorm 则变为

$$\mathrm{RMSNorm}(x')=\frac{(11,12,13)}{\sqrt{\frac{434}{3}+\epsilon}}.$$

当 $\epsilon = 0$ 时，LN 约为 $(-1.224745, 0, 1.224745)$；保留正 ϵ 时应使用上述式. RMSNorm 不具有平移不变性.

6.6 每层 $x \mapsto (1+a)x$，十层导数为 $(1+a)^{10}$. 分别为 $1.1^{10} \approx 2.593742$、$.9^{10} \approx .348678$ 和 0. 即使每层含恒等分支，其他分支仍可放大、衰减或抵消它；加入归一化后，还需分析其雅可比与方向，不能单凭网络图上的直通线保证稳定.

6.7 允许矩阵为

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

前缀位置 1 若预测词元 2、位置 2 若预测词元 3，答案已在可见输入中，形成泄漏. 位置 3 预测首个回答词元 4 时没有直接看到 4，但应按定义仅对回答区域监督，避免把双向条件区当作普通因果训练样本.

6.8 按本章带偏置普通 FFN 及两次归一化约定，编码器为 $4d^2 + 2df + 9d + f$，代入得 7087872；解码器增加一套注意力和一次归一化，为 $8d^2 + 2df + 15d + f = 9451776$. 未计嵌入、位置参数、词表头、堆叠外末端归一化. 无偏置 SwiGLU 等矩阵预算取 $g = \dfrac{2f}{3} = 2048$，只移除原 FFN 的 $f + d = 3840$ 个偏置，得到 7084032、9447936；这里仍保留注意力偏置和归一化参数.

6.9 仍记 $p = \sigma(2\sqrt{2})$、$c = 2p - 1$. 双向权重为 $\left[\begin{smallmatrix} p & 1-p \\ 1-p & p \end{smallmatrix}\right]$，输出 $O = [(-c, c); (c, -c)]$，第一残差为 $U = [(1-c, 3+c); (4+c, 2-c)]$. 两行 LN 仍为 $(-1, 1), (1, -1)$，FFN 输出仍为 $(0, 1), (3, -2)$；故 $H = [(1-c, 4+c); (7+c, -c)]$. 末端二维、$\epsilon$ 为零的 LN 仍给相同符号模式，最终概率和损失恰好不变. 这是退化低维例的遮蔽效应；中间表示已变，第一行通过注意力依赖第二行，不能据末端相同否认泄漏.

6.10 若 $Z = HE^{\mathsf{T}}$，则 $\overline{E}_{\text{out}} = \overline{Z}^{\mathsf{T}} H$.未作为输入的词元仍作为词表候选参与 Softmax，通常有非零分数梯度，故对应 E 行可更新. 重载应核对共享参数关系或等价共享实现、形状和词表 ID、固定位置配置，在评估模式下比较同输入首步 logits 及确定解码首词元；不能仅检查两个矩阵初值数值相等.

6.11 固定权重、输入、精度与评估模式，先关闭 Dropout. 核对词元 ID、嵌入缩放、张量布局以及 QKV 权重的拼接顺序或转置约定，再比较投影结果. 使用短输入核对掩码的真假含义、广播维度和因果位置，随后逐层比较位置变换后的 Q/K、注意力输出、残差、归一化、激活及前馈输出；归一化的前后位置和数值稳定项都应一致. 最后检查词表输出权重是否共享、偏置及 logits. 训练损失还需单独核对标签移位. 中

间量及首步输出在约定容差内一致后，再检查缓存递增位置和完整生成；采样需固定设置，微小数值差异仍可能改变离散生成路径.

第七章　位置表示

7.1

$$\Pi = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad \Pi M \Pi^{\mathsf{T}} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

每列是互异单位向量，故 $\Pi^{\mathsf{T}}\Pi = I$. 变换后的允许位置不再是原编号下的下三角；因果时间顺序不能在保持原掩码时任意置换.

7.2 设 π 是置换，逐行函数满足 $[f(\Pi X)]_i = f(X_{\pi(i)}) = [\Pi f(X)]_i$，前提是各行共享同一个函数且无额外位置依赖. 再求和得到 $\sum_i f(X_{\pi(i)}) = \sum_i f(X_i)$. 前者输出随输入一起重排，称等变；后者输出不变，称不变. 引入位置编码或因果掩码后，应重新审查前提.

7.3 两组频率为 1 与 $\frac{1}{100}$，因此

$$p_0 = (0, 1, 0, 1),$$

$$p_2 = (\sin 2, \cos 2, \sin 0.02, \cos 0.02).$$

由各对的余弦差公式可得内积

$$p_0^{\mathsf{T}} p_2 = \cos 2 + \cos 0.02 \approx 0.583653.$$

每对坐标的平方和为 1，所以任意 t 都有 $\|p_t\|^2 = 2$；范数恒定不代表不同位置正交.

7.4 对 $v(\theta) = (\sin\theta, \cos\theta)^{\mathsf{T}}$，$v(\theta+\delta) = \left[\begin{smallmatrix} \cos\delta & \sin\delta \\ -\sin\delta & \cos\delta \end{smallmatrix}\right] v(\theta)$，即标准旋转 $R(-\delta)$. 标准列向量 RoPE 的 $R(\delta)$ 则为上右负、下左正. 交换 $\sin, \cos$ 坐标次序改变角度方向的矩阵表示，必须连同配对和行列约定转换.

7.5 本题点积为 $\cos\Delta + 7\sin\Delta$. 前三组位置差均为 2，结果均约 5.948935；$(3, 1)$ 的差为 -2，结果 $\cos 2 - 7\sin 2 \approx -6.781229$. 余弦部分为偶函数，正弦交叉项为奇函数，因此反转距离一般改变分数，并非只依赖距离绝对值.

7.6 分块写 $q = (q_R, q_U)$、$k = (k_R, k_U)$，则分数为 $q_R^{\mathsf{T}} R(\Delta\omega) k_R + q_U^{\mathsf{T}} k_U$. 例如 $q = (1, 0, 2, 3)$、$k = (1, 0, 4, 5)$，结果为 $\cos(\Delta\omega) + 23$. 最后两维未旋转，其 23 项不随位置变化；仍需在完整注意力中按定义缩放.

7.7 若新查询正确绝对位置应为 t 而实现设为 0，历史键 j 使用 Rj，则实际角度为 $j\omega$，正确角度为 $(j-t)\omega$，误差为 $t\omega$. 形状和 Softmax 归一化仍可完全合法，但打分

函数已改变. 应以固定完整前向与逐步缓存前向的同位置输出作比较，并同时固定频率配置.

7.8 函数在任意位置可计算只说明表达式有定义，训练未覆盖的相位、干扰数量和长距离任务仍可能失效. 取 $q = k = (1, 0)$，分数为 $\cos(\Delta\omega)$，可随距离先降后升，直接反驳逐样本单调衰减. 讨论泛化须在不同长度、证据位置与干扰密度下评估，并同时检查短文本退化和实际资源成本.

第八章　自回归语言建模

8.1 数据侧移位：输入 (BOS, a, b)，目标 (a, b, EOS)，三个位置有效. 损失侧移位：输入和标签均可提供 $(\mathrm{BOS}, a, b, \mathrm{EOS})$，内部用前三个 logit 对后三个标签，BOS 不作为目标，最后输入位置无后继监督. 若先数据移位再让损失内部移位，BOS 位置被用于预测 b，a 位置预测 EOS，实际变成跨两步目标并丢失边界项.

8.2 概率为 $.7 \times .5 \times .8 = .28$，负对数似然 $-\log .28 \approx 1.272966$，平均三目标损失约.424322. 遗漏 EOS 得到前缀概率.35，表示生成了 a、b 但尚未规定此处结束；这不是同一完整响应事件.

8.3 对 $\tau > 0$，$z_i > z_j$ 当且仅当 $\frac{z_i}{\tau} > \frac{z_j}{\tau}$，指数单调且共享分母，因此排序保持. 若有 m 个并列最大值，其他分数严格较小，$\tau \to 0^+$ 时概率在这 m 项上各趋 $\frac{1}{m}$. 确定贪心按固定规则只输出一项，因此分布极限与确定规则的单点输出不相同.

8.4 $\rho = .5$ 时前一项.4 不足，前两项.7 达到门槛，集合为前两项，概率 $(\frac{4}{7}, \frac{3}{7}, 0, 0)$. $\rho = .95$ 时前三项.9 不足，须保留四项，概率不变. 删除首个跨过门槛的词元会留下累计不足阈值的集合；第一项即跨门槛时甚至可能错误得到空集.

8.5 四条完整路径对数为 $\log .30 \approx -1.203973$、同值、$\log .36 \approx -1.021651$、$\log .04 \approx -3.218876$. 规定并列优先 EOS，完成候选继续占束位. $K = 1$ 先保留 a，再保留 a-EOS；$K = 2$ 先保留 a、b，再保留 b-EOS 与 a-EOS，两者都完成，最优为 b-EOS. 若并列选 a-c，下一步必 EOS，得分仍不变；明确并列规则后可复现.

8.6 每个后继对数概率不大于 0，所以任何扩展分数 $s' \leqslant s$，原前缀分数是其后代原始分数上界. 若用平均对数分数，前缀长度 1、分数 -2，追加概率 $e^{-.1}$ 后得 -2.1，平均由 -2 升至 -1.05. 故不能把原始分数的单调上界直接用于长度归一化剪枝.

8.7 把 EOS 一步计入 T，$P(T = t) = (1-\epsilon)^{t-1}\epsilon$，$t \geqslant 1$. 由尾和公式 $\mathbb{E}T = \sum_{t \geqslant 0} P(T > t) = \sum_{t \geqslant 0} (1-\epsilon)^t = \frac{1}{\epsilon}$；非 EOS 词元数期望为 $\frac{1-\epsilon}{\epsilon}$. 永久截断 EOS 会使实际采样分布的终止概率为零，必须依赖长度上限等外部条件停止，这不再是同一分布.

8.8 为每个请求保存独立完成标记，完成后不再采样有效词元或增加有效长度；若静态批次仍需占位，填充输出不得成为后续真实条件. 动态压缩批次时同步重排请求 ID、缓存块和位置. 继续生成再裁文本会多用计算、改变随机数消费和缓存长度，甚至触发工具调用；输出裁剪不能替代执行停止.

8.9 停止串可取 `END`，相邻两个解码片段分别为 `E` 与 `ND`，各自都不等于完整停止串. 需要在累积解码流中保留足够长的后缀并与新片段共同匹配；字节级词元还应处理不完整字符的增量解码状态. 命中后按协议决定是否保留停止串，且匹配策略须与模型 EOS 分开记录.

8.10 有限数据与模型误差使生成前缀偏离训练支持，此为训练与访问分布错配的一种来源；截断或极低温度还可能放大重复、删去正确候选，这是解码分布变化；EOS 被屏蔽或最大长度过短则属终止协议错误. 教师强制正确实现最大似然，并不保证有限模型无误差；应分别用前缀扰动、固定模型解码对比和停止回归定位，而非统一归咎一种机制.

8.11 “2023 年”与证据给出的 2024 年冲突，既是相对于该证据的上下文矛盾；若 2024 年也是目标范围内的有效事实，它同时构成事实错误. “由甲”没有证据支持，属于无依据生成；引用证据包中不存在的条目属于虚构归因. 降低温度只把概率集中到原有高分路径，若这些声明原本得分最高，错误会更稳定. 控制可包括限定只能依据证据作答、验证声明到来源的映射、对年份和实体做规则或工具核验，以及证据不足时删除该声明或拒答.

第九章　模型评估

9.1 $P=\dfrac{TP}{TP+FP}, R=\dfrac{TP}{TP+FN}$ 代入调和均值得 $F_1=\dfrac{2TP}{2TP+FP+FN}$，分母为零时另定口径. 1000 例中正类 10，模型全判负，准确率 99%，正类召回 0，F1 可按约定为 0. 高准确率主要来自负类占比，不能证明识别稀少正类.

9.2 每个错误单标签预测给真实类贡献一个 FN、预测类贡献一个 FP，正确预测贡献一个 TP. 故总 TP 为正确数 c，总 FP 与 FN 均为 $n-c$，Micro-F1 为 $\dfrac{2c}{[2c+2(n-c)]}=\dfrac{c}{n}$. 多标签例中两个样本真值均 $\{A,B\}$、预测均 $\{A\}$，Micro-F1 为 $\dfrac{4}{4+2}=\dfrac{2}{3}$，严格样本准确率却为 0. 需固定准确率具体指子集准确还是逐标签准确.

9.3 记 C_{ay} 为采取 a、真值 y 时成本，p 为真阳概率. 行动 1 风险 $(1-p)C_{10}+pC_{11}$，行动 0 风险 $(1-p)C_{00}+pC_{01}$. 若 $D=C_{10}-C_{00}+C_{01}-C_{11}>0$，选 1 要求 $p>\dfrac{C_{10}-C_{00}}{D}$；D 非正时须直接比较或反向不等式. 增加弃权行动 a=r 时，只要其条件风险低于两者即可弃权，风险成本与后续人工能力必须定义.

9.4 按分数由高到低取标签 $(0,1,1,0)$. 从包含全四项开始提高阈值，精确率依次 $\frac{2}{4},\frac{2}{3},\frac{1}{2},\frac{0}{1}$，先升后降. 召回率为已纳入正例数除固定总正例数，去掉低分项只可能减少或保持分子，故随提高阈值不增. 阈值变化若同时改变数据总体或任务判定，该单调证明不再适用.

9.5 设同一串文本的总概率为 p，一种切成 2 词元，另一种切成 4 词元，即便对该字符串赋相同总概率，PPL 分别 $p^{-\frac{1}{2}}$ 与 $p^{-\frac{1}{4}}$，单位不同不能按小者判断更强. 只计回答时定义目标位置掩码 m_t 并计算 $\exp[-\dfrac{\sum m_t \log p(y_t \mid \text{合法前缀})}{\sum m_t}]$，EOS 是否计入预先固定，提示可作上下文而不作目标. 空回答不能以零分母计算.

9.6 明确把候选“甲乙戊”追加一个“甲”，参考仍“甲乙丙丁”. 多重集合截断后重合仍 2，候选长 4、参考长 4，Token F1 为 $\frac{4}{8}=0.5$；一元截断精确率 $\frac{2}{4}=0.5$. 若只按候选词是否出现于参考而不截断，两次甲都被计为匹配，错误得到 $\frac{3}{4}$，鼓励重复常见词. 若重复的具体词不同，须按实际计数重算，不能只加长度.

9.7 n 候选均匀无放回选 k，共 $\binom{n}{k}$ 个子集，全失败子集 $\binom{n-c}{k}$，所以至少一通过的子集比例为 $1-\dfrac{\binom{n-c}{k}}{\binom{n}{k}}$，约定 $k>n-c$ 时失败项为 0. 作为固定候选集的组合恒等式不要求候选独立；将其解释为未来 k 次同分布尝试覆盖率的无偏估计则需相应抽样假设. 即使覆盖成功，产品还要从候选中选择正确者，完美验证选择不是组合公式自动保证的.

9.8 对同一答案对交换 A/B 呈现顺序，另用保真短版/长版交叉展示，再比较裁判与候选是否同家族；随机分配并盲去模型标识，形成正交或平衡设计. 按真值、长度和家族分别估计偏差并检查交互，固定裁判协议. 总体一致率可被容易样本主导，掩盖少数高风险或同家族偏好，需分层人工锚与配对差.

9.9 从 $(\hat{p}-p)^2 \leqslant \dfrac{z^2p(1-p)}{n}$ 移项，得 $(n+z^2)p^2-(2n\hat{p}+z^2)p+n\hat{p}^2 \leqslant 0$，两根即 Wilson 端点. 全失败 $\hat{p}=0$ 时通过率上界 $\dfrac{z^2}{n+z^2}$；如 n=20、z=1.96，为 0.1611. 零失败的失败率单侧 95% 精确上界为 $1-0.05^{\frac{1}{n}}$，n=20 时约 0.1391. 两者置信方向与构造不同，不能比较数字后声称哪个必然更准确.

9.10 用户 A 发 1 请求且成功，B 发 9 请求全失败. 请求平均 $\frac{1}{10}=0.1$，用户平均 $\frac{1+0}{2}=0.5$；若交换各用户的成功情况，请求平均 0.9 而用户平均仍 0.5. 前者估计随机请求体验，后者估计随机用户平均体验. 按用户重采样后仍需使用与目标对应的聚合权重，抽样单位与计分权重是两件事.

9.11 固定 100 题运行稳定性关注在这些题上改变随机种子的变化，需种子/运行可重复、系统配置固定并说明共享状态. 推断新题总体需题目代表目标分布，按题或来

源簇重采样；同题五次相关，不能当 500 独立题. 若想同时推断题和运行变化，可用交叉/分层设计并保留两层结果；仅对五个总分计算区间不覆盖题集代表性误差.

9.12 每天窥视并首次显著即停等于多次选择，固定样本 p 值的单次错误率不控制整条路径. 可预先固定终止样本量与一次主检验；若必须分阶段决策，预定观察时点与 alpha 分配，或使用有明确假设的序贯检验/置信序列. 固定主要指标及排除规则，不能看到结果后临时改变停止和显著性标准.

9.13 可行域如质量 $Q \geqslant 0.8$、高损失错误率 $e \leqslant 0.01$、p95 延迟 $T \leqslant 5$ 秒、成本 $c \leqslant 0.05$ 元/请求. 令 A 为 $(Q, T, c) = (0.82, 1, 0.01)$，B 为 $(0.88, 2, 0.02)$，C 为 $(0.93, 4, 0.04)$，且均满足风险约束，三者无一在全部目标上不差且至少一项更好. 先过硬约束，再按预先商定业务效用选择前沿点；权重不能依据最终测试结果事后调至偏爱某候选.

9.14 可支持给定测试混合上总体点估计提高，不能据少量高风险样本声称其错误率足够低或已安全. 报告该切片样本量和区间，增加按危险机制分层的独立样本及真实失败附近的受控变体，同时保留正常样本测误拒. 富集采样用于风险机制检验，若估总体率须按真实基率加权，不能把富集数据比例当生产基率.

9.15 Astra 行为为 v2026.08.08 离线子集部分得分；Fable 为作者 8 月任务版及后续任务/脚本修复，108 任务、1080p、500 步、max effort、五次运行，需模型评分任务用 Opus 4.8，且安全干预计零. Astra 采用其发布配置，不能假定相同集合、框架、裁判、预算或安全回退，故 72.6 与 77.9 不能直接证明模型能力差. 受控实验应冻结同一合法任务快照、环境、验收器、工具、预算和安全配置，两系统按同题重复运行并保存逐例结果，以任务为配对单位，同时报告 partial 与 strict、成本及未知失败.

9.16 只有一提示且五约束四通过时，指令级为 $\frac{4}{5}$，提示级因非全通过为 0. 严格与宽松评分另按 IFEval 规定的有限输出转换判定，不能任意把失败约束忽略. 格式、长度等规则不检测全部事实，故还须对业务论断、引用和授权作独立验收，不能把 80% 规则通过率称为 80% 事实准确率.

9.17 基础层可用 C-Eval/CMMLU 检查中文知识与回归，但不涵盖企业私有事实、权限和时效. 检索层用企业保留查询的 Recall、nDCG 与无证据误返回，仍不证明生成忠实；端到端按正确、引用支持、恰当拒答、权限违规、时延成本计分. 保留集来自经授权脱敏的真实请求、专家构造的版本冲突和不可回答案例，并按文档/客户及时间隔离；禁止把已用于调提示的样本重新称测试集.

第十章　模型安全

10.1 问答模型的资产含输入文档、隐私、回答可信度和来源；不可接受后果包括泄露或高损失误导. 增加写权限后还涉及账户、库存、订单与业务状态，新增模型提案

到授权网关、网关到业务系统、批准到执行的信任边界. 控制需从内容检查扩展至对象权限、参数约束、幂等与回执核验，不能仅提高拒答率.

10.2 互斥完备场景 S_i 下，$E[L] = \sum_i P(S_i)E[L \mid S_i]$. 若定义可加损失分量 $L = \sum_j L_j$，线性期望给 $E[L] = \sum_j E[L_j]$，不需独立. 若两条攻击路径均导致同一次 100 单位泄露且各概率 0.1、完全重叠，分别算 0.1 乘 100 再相加得到 20 而真实为 10；路径不是互斥场景，也不是两个不同损失分量，须去重或按最终事件划分.

10.3 设风险 A 等级 $(1,4)$、B$(2,2)$，编码 1、2、3、4 时乘积均 4；采用保持顺序的编码 $(1,3,4,5)$ 时 A 为 5、B 为 9，B 高；编码 $(1,2,3,10)$ 时 A 为 10、B 为 4，A 高. 等级只规定顺序，任意单调重标度可改变乘积关系，因此不能解释为概率乘真实损失或风险倍数.

10.4 设漏放事件 A、B，Fréchet 界为 $\max(0, 0.2+0.3-1) \leqslant P(A \cap B) \leqslant \min(0.2, 0.3)$，即 0 至 0.2. 独立时为 0.06；A 包含于 B 时为 0.2；两者互斥时为 0. 不能默认层数增加就乘法降低风险，应测条件漏放 $P(B \mid A)$ 和共同失败原因.

10.5 模型行为控制可减少建议外发的概率；授权控制检查任务目的、收件对象和数据级别，阻止即便格式合法的越权发送；短期最小凭据限制可触达资源和端点；渲染控制防止把工具文本中的主动资源或链接自动执行/加载. 它们保护不同路径，内容分类器不能代替确定性授权，工具成功也不证明数据允许披露.

10.6 防御例可用对象别名或大小写、编码差异：检查时把别名视为公开对象，解析器最终映射到受限对象. 应先用唯一规定规则规范化并解析为稳定资源 ID，再按该 ID 和当前租户授权，执行时使用同一解析结果及版本. 若重定向或资源映射可变化，重新校验最终对象；拒绝歧义输入，不维护两套不一致的路径解释.

10.7 摘要字段包括主体/租户、动作类型、工具版本、最终资源 ID、规范参数、目的地址、数据级别、对象/权限版本和有效期. 参数或资源版本变化后重算摘要并检查批准绑定；权限撤销即使摘要不变也应阻止执行. 批准只在其明确范围及期限内有效，不是给予模型后续任意操作的令牌.

10.8 账本按租户和稳定幂等键唯一认领，绑定动作摘要，状态记录发送中、未知、确认成功或明确失败. 超时后按原键查询工具结果，确认成功则补回执；不能确认则保持未知并人工对账. 新键使服务端把同一意图当新动作，可能再次扣款或发信. 若远端没有幂等或可靠查询，只能停止自动重发，不能由本地账本宣称恰好一次.

10.9 哈希一致说明与给定字节基准相同，前提是基准可信；签名说明受信密钥对某制品背书，但不保证行为安全；加载成功说明格式与部分环境兼容；行为评测通过只支持指定样本和协议下的观察. 四者均不独立证明生产安全，需要制品身份、可信来源、隔离加载和任务回归的联合证据.

10.10 谱系可为样本 ID 到清洗/切片修订、训练快照、优化运行、模型/适配器哈希、索引/缓存、部署修订. 删除原文后还需检查副本、备份、向量、摘要、日志、评测集及已训练参数等派生资产；不同资产删除语义不同. 追踪用于识别影响范围，不能仅删除一条数据库记录就声称模型已遗忘该信息.

10.11 题目未指定参数，取危险召回 $r = 0.9$、正常误拦 $f = 0.02$，基率 π 时拦截精确率为 $\dfrac{r\pi}{[r\pi + f(1-\pi)]}$. $\pi = 0.01$ 得 $\dfrac{0.009}{0.0288} = 31.25\%$；$\pi = 0.1$ 得 $\dfrac{0.09}{0.108} = 83.33\%$；$\pi = 0.001$ 仅约 4.31%. 相同过滤器在低基率下大多数拦截可能是假阳性，审核负担应按请求量与误拦率共同估计.

10.12 若独立同分布失败概率 q，n 次零失败概率 $(1-q)^n$；令其等于 α 反解单侧上界 $q_U = 1 - \alpha^{\frac{1}{n}}$. 要求 95% 上界严格小于 0.01，须 $n > \dfrac{\log 0.05}{\log 0.99} \approx 298.073$，最小整数 299. 零失败不代表 q 为 0，相关样本或变化分布会破坏该二项界的前提.

10.13 每语言分别准备正常、风险及分层场景；漏放分母为该语言真实危险请求数，误拦分母为真实正常请求数，控制覆盖分母为预先列出的应受控路径或场景数并注明加权方式. 若某切片只有危险样本，误拦率未定义，不得填 0；只有正常样本则危险漏放未定义. 总体聚合须按目标分布权重，不能用多语言题量代替真实流量比例.

10.14 把 10:01 成功改为发送后超时，则 10:03 告警不能证明已泄露，也不能证明未执行. 先停用发送并撤销凭据限制扩大，同时保全动作键、收件信息及工具端日志；按原键查询和人工对账确认影响. 检测时延若无已确认执行时刻应报告从首次可疑行为或告警起的明确替代口径，不能继续写执行到告警 2 分钟. 恢复前先处理未知动作，防止恢复后盲重放造成真正泄露.

第十一章　上下文学习

11.1 上下文学习在固定 θ 下改变条件，比较 $p_\theta(y \mid x, D_1)$ 与 $p_\theta(y \mid x, D_2)$；微调更新 $\theta' = \theta - \eta \nabla_\theta \mathcal{L}$，之后即使不再提供 D 也可能改变输出. 零样本只说明当前上下文没有任务示例，不排除预训练中已见任务或同源数据. 两者均须隔离评测样本.

11.2 可令 $F(S) = \sum_{i \in S} r_i + \lambda \sum_{k} \min(1, \sum_{i \in S} a_{ik}) - \gamma \sum_{i<j \in S} \mathrm{sim}(i,j)$，约束 $\sum_{i \in S} l_i \leqslant B_D$. r_i 来自查询相似度，a_{ik} 来自预先标注技能，冗余来自示例内容；权重在开发集确定. 禁止使用测试答案、测试判定结果或按最终测试分数逐题挑示例，查询本身可作为已允许输入.

11.3 选同一对示例 $(x_a, A), (x_b, B)$，顺序一为 a 后 b，顺序二为 b 后 a；最后标签分别为 B 和 A. 对每个保留问题同时运行两序列，固定模板、词元预算、模型和种子方案，记录配对差；在多组示例对上重复. 这样内容集合不变，差异可归于顺序及其位置交互；若同时替换示例，则无法隔离二者.

11.4 按逐类相除，未归一化量为 $(\frac{5}{6},\frac{3}{2},1)$，和为 $\frac{10}{3}$，故校准后的概率为

$$(\frac{1}{4},\frac{9}{20},\frac{3}{10})=(0.25,0.45,0.30).$$

原最大类变成第二类. 除去内容无关偏好依赖其估计有效；真实业务标签可能本就不均衡，强迫均衡会抹掉有用先验，应使用开发数据和校准目标论证.

11.5 令五路径概率为 $0.30,0.25,0.20,0.15,0.10$，首条答案 A，第二、三、五条答案 B，第四条答案 C. 最大概率路径给 A；答案边缘概率为 $P(A)=0.30,P(B)=0.55,P(C)=0.15$，故边缘最大给 B. 概率必须覆盖同一分布并和为 1，不能只列未归一化分数冒充路径概率.

11.6 令 $Z_k=\mathbf{1}\{Y_k=a\}$，独立同分布时 $E[Z_k]=p_a$、$\operatorname{Var}Z_k=p_a(1-p_a)$. $\hat{p}_a=K^{-1}\sum Z_k$ 因此无偏，方差为 $\frac{p_a(1-p_a)}{K}$；相关时另有 $2K^{-2}\sum_{i<j}\operatorname{Cov}(Z_i,Z_j)$. 温度及截断改变实际采样分布，频率估计的是变换后的答案边缘，而非未变换模型分布.

11.7 二元多数需要至少三票，概率为 $10(0.4)^3(0.6)^2+5(0.4)^4(0.6)+(0.4)^5=0.31744$，低于单次 0.4. 三答案可取正确 A 概率 0.4，错误 B、C 各 0.3；A 虽不过半却是唯一众数. 大样本最高票收敛到众数需要独立同分布，不能套二元超过半数公式.

11.8 奇数只防止两类别平分全部选票，多类别仍可出现两个最高各 2 票. 预先规定先比较独立验证器，再在并列者中用固定规则或额外预算采样；若必须稳定可用规范化答案的固定顺序，但应说明这种规则不增加正确性. 禁止查看标准答案再解平票；解析失败须另计.

11.9 独立同分布、单次正确概率 p 时，候选至少一个正确为 $1-(1-p)^K$；它只描述可用候选覆盖. 最高票还取决于错误答案如何分散及平票规则. 验证选择又取决于验证器漏判、误判和选择政策，理想无误选择器下成功率才可等于覆盖率. 相关或自适应生成时第一式本身也未必成立.

11.10 当前领先差 $5-2=3$. 只剩 2 条时第二名至多 4，原第一名不可被追平，可停止而不改变最高票结果；仍须满足其他强制验收. 剩 4 条时第二名可到 6，不能提前确定. 若规则对平票的处理不同，安全条件仍可采用严格领先差大于剩余候选数.

11.11 错误候选可执行表达式 $30(6-1)=150$，在当前参数上数值碰巧正确，却把满减 30 解释为赠送一件；换为单价 40 仍满减 30 时，该解释给 200，正确为 210. 更直接的错误答案 $30\times6=180$ 算术可执行但遗漏优惠. 验证器需核对实体、单位、数量、优惠条件和运算与题意的对应，并用受控变体检验泛化；只检查程序能运行不足够.

11.12 无缓存时增加示例长度 ΔL 会被 K 候选重复读取，附加输入成本近似 $K\Delta L$；增加候选还需完整提示与新输出. 共享前缀可降低重复预填充/计费，但缓存读取仍

非零，且身份必须相同. 并行降低部分墙钟等待，不自动降低词元或设备总成本，还受并发容量限制. 用开发集估计各选择的质量边际提升，在同费用和期限下求可行组合，不凭候选数决定.

11.13 例如长度答案统一为米，允许有限数值容差；集合按无序集合规范化，未提供单位的值只有任务明确默认单位时才转换. 空字段、无答案和解析失败用不同状态. 若 K 条中 m 条可解析、其中 n 条支持答案 a，总体支持率为 $\frac{n}{K}$，条件支持率为 $\frac{n}{m}$；m 为零时后者未定义. 排除失败会夸大可交付支持.

11.14 报告模型、模板、示例身份与顺序、分词长度、采样、解析、工具和预算；新模型需重测格式、校准、任务正确性和成本，旧结论不能按模型名继承. 同题 K 候选属于嵌套重复，先按题得到固定决策或以题为簇重采样. 报告问题数与每题候选数，不将 100 题五候选称为 500 个独立测试任务.

第十二章　信息检索

12.1 片段记录包含 `document_id`、revision、`chunk_id`、页码/表行列、标题路径、有效日期、来源哈希、权限域和提取器版本. 表格值还须携带表头、单位、脚注及适用对象，避免把“10”脱离“万元/季度”. 引用定位须能复原原版本，撤权过滤发生在检索与返回两阶段；仅存文本向量不足以证明来源.

12.2 由 Bayes 及条件独立，后验对数赔率等于先验对数赔率加 $\sum_t [x_t \log(\frac{p_t}{u_t}) + (1 - x_t) \log(\frac{1-p_t}{1-u_t})]$. 将与文档无关项合并后，出现词的系数是 $\log[\frac{p_t(1-u_t)}{u_t(1-p_t)}]$. BM25 还需估计概率、平滑 IDF、从二元出现推广到词频饱和，并引入长度归一化；这些不是上述独立假设自动推出的唯一形式.

12.3 令 $K = k_1(1 - b + b\frac{\ell}{\bar{\ell}}) > 0$，词频因子 $f(t) = \frac{(k_1+1)t}{t+K}$. 有 $f'(t) = \frac{(k_1+1)K}{(t+K)^2} > 0$、$f''(t) = -\frac{2(k_1+1)K}{(t+K)^3} < 0$，即递增且边际收益递减. 若也考察长度，$\frac{\partial f}{\partial \ell} = -\frac{(k_1+1)tk_1 b}{[\bar{\ell}(t+K)^2]}$. b 为 0 不校正长度，b 为 1 采用全长度比例；极短或极长文本的效果仍取决于词频与语料.

12.4 在冻结 $N, n_t, \bar{\ell} = 100$ 的评分快照下，把 B 长度改成 100，长度项变 1.2；分数为 $\log(\frac{101}{10.5})(\frac{6.6}{4.2}) + \log 2(\frac{4.4}{3.2}) \approx 4.5104 > 3.8058$. 可从 B 删除无关词使目标词频仍为 $(3, 2)$. 若真的修改整个语料并重建，平均长度等统计也会变；题目没有给其余文档，不能声称冻结统计与重建统计完全相同.

12.5 点积分数 $s_j = u^T v_j$，$\mathcal{L} = -\frac{s_+}{\tau} + \log \sum_j e^{\frac{s_j}{\tau}}$，故 $\nabla_u \mathcal{L} = \frac{\sum_j p_j v_j - v_+}{\tau}$. 高分假

负例具有较大 p_j，梯度下降将查询推离它，损害真实相关邻域. 应识别多正例、控制负例污染；不能仅以训练损失下降断言召回更好.

12.6 取查询 $q=(1,0)$，文档 $a=(10,10), b=(0.9,0)$. 内积 10 与 0.9 使 a 优先；余弦 $\frac{1}{\sqrt{2}}$ 与 1 使 b 优先；到 q 的距离分别 $\sqrt{181}$ 与 0.1 也使 b 优先. 若都单位化，余弦和内积等价且平方欧氏为 $2-2q^Td$，上述分歧依赖范数未统一.

12.7 更宽探索可访问更多候选、减少局部贪心遗漏，但有限预算和图连通性仍限制结果. 权限过滤可能断开原图或消除近邻桥梁，先截断再过滤也可能返回不足 K 项. 应与同权限集合内的精确搜索比较，必要时过滤感知搜索或小集合精确回退；探索宽度不是普适精确性证明.

12.8 每子空间需 $\log_2 256=8$ 位，48 段共 48 字节. 每子向量维数 $\frac{768}{48}=16$，全部码本有 $48\times256\times16=196608$ 个中心元素；若每元素 4 字节，另为 786432 字节共享码本. 每对象的原向量、ID 和索引连接没有计入 48 字节.

12.9 令重建误差 $e=\hat{x}-x$，平方距离差为 $\|q-\hat{x}\|^2-\|q-x\|^2=-2(q-x)^Te+\|e\|^2$，绝对值不超过 $2\|q-x\|\|e\|+\|e\|^2$. 若 A 和 B 各有误差界 ϵ_A,ϵ_B，原距离间隔大于二者和可保持排序，否则不能保证. 非平方距离还满足反三角不等式界 $|\|q-\hat{x}\|-\|q-x\||\leqslant\|e\|$.

12.10 两路为 (A,B,C) 及 (B,D,A). c 为 1 时 A 为 $\frac{1}{2}+\frac{1}{4}=0.75$，B 为 $\frac{1}{3}+\frac{1}{2}=\frac{5}{6}$，C 为 $\frac{1}{4}$，D 为 $\frac{1}{3}$；c 为 60 时 A 为 $\frac{1}{61}+\frac{1}{63}\approx0.03227$，B 为 $\frac{1}{62}+\frac{1}{61}\approx0.03252$，C 为 $\frac{1}{63}\approx0.01587$，D 为 $\frac{1}{62}\approx0.01613$. 均为 B、A、D、C，但顶部 B-A 差从 $\frac{1}{12}$ 降至约 0.000256，c 越大越弱化顶部位次差异.

12.11 重排只为候选集合内对象评分，唯一证据不在集合中时不存在可提升对象. 加深召回增加检索 I/O、重排模型输入对数、上下文长度和延迟；联合编码器尤其不能把成本视为常数. 应同时绘制候选覆盖与预算曲线，定位是召回遗漏还是排序失误.

12.12 权威权限/墓碑先提交，查询入口与返回前强制过滤，使旧索引暂未删除也不可见；随后发送版本化删除事件更新倒排、图节点和向量缓存，失效依赖该文档的检索缓存. 重复删除幂等，乱序旧写入不能覆盖墓碑；物理压缩可异步但须有完成水位. 负缓存也依赖语料/权限版本，不能仅追踪已命中文档.

12.13 沿用全部相关等级 $3,2,1$，新排名 $(1,0,3,2)$ 的 DCG 为 $1+\frac{7}{2}+\frac{3}{\log_2 5}\approx5.7920$；IDCG 仍 $7+\frac{3}{\log_2 3}+\frac{1}{2}\approx9.3928$，nDCG 约 0.6166. 此时三个相关对象均在前四，Recall 为 1，首条相关所以 RR 为 1；原例为 Recall$\frac{2}{3}$、RR$\frac{1}{2}$、nDCG0.5225. 三个量分别关注覆盖、首次命中和等级排序.

12.14 设精确向量前 10 均为措辞相近却版本错误的文件，ANN 全部复现，ANN Recall@10=1；唯一正确版本排名 100，任务 Recall@10=0. 此时首先改表示、版本过

滤或查询协议，而不是只加图搜索宽度．若精确前 10 已有证据而 ANN 漏掉，才是索引近似误差主导．

第十三章　检索增强生成

13.1　序列潜变量 z 在整条答案中固定，由全概率得 $p(y \mid q)=\sum_{z} p(z \mid q)\prod_{t} p(y_t \mid q, z, y_{<t})$．逐词元潜变量 z_t 分别边缘化则为 $\prod_{t}\sum_{z_t} p(z_t \mid q)p(y_t \mid q, z_t, y_{<t})$．展开后一式包含 $z_1 \neq z_2$ 等跨文档项，前式不含，通常不能交换．它们是不同生成假设，不是同一实现的两种加速写法．

13.2　取权重 $(0.5, 0.3, 0.2)$，两位置条件概率分别 $(0.8, 0.2), (0.2, 0.8), (0.5, 0.5)$，均针对同一目标前缀．序列概率 $0.5(0.16)+0.3(0.16)+0.2(0.25)=0.178$；词元概率 $(0.4+0.06+0.1)(0.1+0.24+0.1)=0.2464$．只留前二且保留原权重，分别为 0.128、$0.46\times0.34=0.1564$，是截断质量而非完整归一化分布．局部重归一化权重 $(0.625, 0.375)$ 后为 0.16、$0.575\times0.425=0.244375$，改变了定义；两种截断不能混报．

13.3　令 $Z=\sum_{i} p_i a_i$ 且固定 a_i 及候选集合，$\dfrac{\partial Z}{\partial s_j}=\sum_{i} a_i p_i(\delta_{ij}-p_j)=p_j a_j - p_j Z$．于是 $\dfrac{\partial(-\log Z)}{\partial s_j}=p_j-\dfrac{p_j a_j}{Z}=p_j-\gamma_j$．候选外文档无此局部梯度；若有效证据完全遗漏，更新最多在错误候选中重新分配概率，不能直接创造候选或修复事实性．

13.4　令 $\Delta=s^{+}-s^{-}$，$\ell=\log(1+e^{-\Delta})$，则 $\dfrac{\partial \ell}{\partial s^{+}}=-\dfrac{1}{1+e^{\Delta}}$、$\dfrac{\partial \ell}{\partial s^{-}}=\dfrac{1}{1+e^{\Delta}}$．梯度下降提升正例相对分数．训练期望取决于采到哪些查询和负例，曝光偏差及假负例会改变优化对象；部署候选与训练候选不同也可能失配．

13.5　d_2 增至 220，$d_2+d_3=340$ 超预算；$d_1+d_3=340$ 也不可行，三片段最短组合仍超 320．最多覆盖两要点：d_2 单独覆盖 A、B 用 220，d_3+d_4 覆盖 B、C 用 220，d_1+d_4 覆盖 A、B 用 320，d_2+d_4 用 320 仍只 A、B．可预先规定先最大覆盖、再最短、再总重排分数，则选 d_3+d_4（1.45 高于 0.85），并明确缺 A；若 A 为必备硬约束，则选 d_2．题设没有要点优先级，不能声称唯一最优．

13.6　原问“它在去年生效时的退费条件是什么”，先解析“它”为产品 P、去年为明确年份 Y．改写一为“产品 P，Y 年生效版本，退费适用条件”；改写二为“P Y 年条款：退款例外与生效日期”．产品、年份、所问条件均不能改变．改成“P 最新退款优惠”可能命中新页面更多，却把历史约束和退费规则变为当前营销内容．

13.7　节点一按目标历史时点读取项目负责人，输出人 ID 及任期；节点二按同一快照查其服务归属区间；节点三按服务 ID 与目标月统计故障，按事件 ID 去重．任务图为一到二到三，服务间计数可在二后并行．负责人变更用有效时间关联而不是当前姓名，所有结果带快照/水位和时间窗；无法得到一致历史快照时说明估计范围．

13.8 来源 ID d7 真实定位到“普通任务上限 30 分钟”，答案却写“维护窗口也一律 30 分钟 [d7]”. ID 格式检查通过，版本与页码定位通过，但来源不包含维护窗口规则，语义支持检查失败. 应返回不支持/需补充例外证据，不能把真实 URL 当作每条论断正确的充分条件.

13.9 回答风险为 $pC_c + (1-p)C_w$，拒答风险为 C_r. 若 $C_w > C_c$，回答条件为 $p > \dfrac{C_w - C_r}{C_w - C_c}$，等号按预定平局规则；阈值超出 $[0,1]$ 时可能总拒或总答. 若 $C_w = C_c$，只比较常数；若大小逆转，不等号方向需重推. 部分回答包含多个论断及不同错误成本，单一二元 p 和一个拒答成本不能表达覆盖与风险结构.

13.10 更新文档提升 revision 并失效引用该版本的证据/答案；撤权先使授权过滤生效，再失效权限域相关缓存；新文档加入提升语料快照或索引水位，使旧无结果缓存重新检索. 负缓存没有已引用文档列表，故仅依赖列表校验无法感知新证据. 缓存键至少绑定查询、模型/表示、语料与权限身份，失效传播未完成时通过权威版本再校验保护可见性.

13.11 以 $(source, object, revision, event_type)$ 作为事件幂等身份，同对象只允许更高版本覆盖，墓碑参与排序. 没有全局版本时可用每对象单调修订或来源游标；这能保证对象内旧写不覆盖新写，却不能保证跨对象同一时点快照. 仅有不可靠时间戳时需回源核对或记录冲突，不能用到达先后冒充源事实顺序.

13.12 取 $P(R) = 0.9, P(S \mid R) = 0.8, P(A \mid R, S) = 0.85$，完整支持路径概率为 0.612，无需假定三事件独立. 另有 0.1 样本未召回证据，其中一半凭参数记忆答对，可使答案正确率至少再增 0.05，但这些答案不具备该证据路径的可追溯支持. 质量报告应分别给事实正确和来源支持，不能把两者相等当公理.

13.13 按可回答且授权、不可回答、无权、上游异常四类保留独立标签. 正确回答率可用可回答且授权题数为分母，过度拒答率用同一分母；不可回答正确拒答率以真正缺证据题数为分母，权限违规率以无权题数为分母. 故障率以全部被评估请求为分母，另列各类故障，不将超时混入正常拒答. 端到端总体成功另按预定四类正确行为聚合.

13.14 比较旧/新重排器，固定初召集合、证据预算、生成器、提示和问题. 分别将证据替换为人工充分包、删除关键证据、加入同实体冲突版本；同时保留正常条件. 若充分包消除差异，瓶颈主要在检索/选择；若删证据仍高分，应审查泄漏和参数记忆. 按同题配对控制种子或重复采样，以题为独立单位；不能给新重排器额外上下文后归因于排序算法.

第十四章　领域应用

14.1　可定义“在授权知识范围内处理一条退换货咨询”：输入订单身份、用户问题及规则版本；输出附证据答复或需人工原因；结束为用户所需信息已核验给出或明确转交. 记录每请求处理时间、首次解决、错误承诺和人工干预，超时也是失败状态. 满意度可附带，但不是唯一结束判据.

14.2　证据记录包含合同 ID 与哈希、修订、生效日、页码及条款位置、原文跨度、结构化字段、单位与例外、提取器版本和核验人. 文件名可以重复或内容被替换，不能固定证据对象. 把“自动续期”提取为布尔值时还须保存期限、通知条件及例外，否则结构虽正确却丢失法律语义.

14.3　至少包含租户/查询者、SKU 及仓库、数量单位、在库/可用/冻结等状态、业务时间和快照版本. 缺身份可查到其他租户；缺单位把箱数当件数；缺状态把冻结库存当可用；缺时间把昨日余额当当前承诺. 涉及实际预留还需原子条件写，查询正确不保证稍后库存仍足.

14.4　新期望时间为 $5+4+0.3\times 8=11.4$ 分钟，净节省 $12-11.4=0.6$ 分钟/任务，即 5%. 100 任务只节省 60 分钟，维护、评估和系统成本未计；若这些摊销超过 0.6 分钟等价成本则净收益可变负. 不能仅比较原 12 与辅助 5 得 7 分钟节省.

14.5　令自动成本 a、正确概率 p、错损 L；人工成本 h、残余错率 e，暂设错损相同且正确损失为 0. 自动风险 $a+(1-p)L$，人工风险 $h+eL$；自动优于人工要求 $p>1-e+\dfrac{a-h}{L}$，L 须正，阈值超范围时作总选某行动解释. 若人工错误也依案例变化，使用 $e(x)$ 并逐例比较；拒答等第三动作须一并纳入.

14.6　观察是“14:02 温度传感器读数超阈值”；关联是“历史同类告警常伴随冷却流量下降”；假设是“可能堵塞导致冷却不足”；确认原因需现场检查、控制实验或维护证据证明具体堵塞并排除替代解释. 模型可组织证据和提出假设，不能将相关性自动升级为已确认故障，更不能据此绕过工业控制许可.

14.7　满意度受流畅、迎合和即时情绪影响，学习效果要观察知识或迁移技能的变化. 可按学生随机分组或分层，使用等价前测后测、延迟保持测试和未见过的迁移题，控制教师与学习时长，按学生为独立单位. 需要报告原始水平、缺失和帮助次数，不能把多轮点赞当独立学习增益.

14.8　以修复金额舍入为例：编译仅证明语法和类型满足编译器；单元/集成测试检查有限输入及接口；评审检查设计、边界和可维护性但仍有漏检；部署验证在真实依赖和流量中是否满足指标. 每层应保存对应证据，测试通过不代表已发布，发布成功也不自动证明所有金额正确.

14.9　简单任务正确率 0.95、困难 0.50 固定. 上线前各占一半，总体 0.725；上线后

简单占 80%，总体 0.86．总分上升 0.135 而各切片能力未变．应以固定任务分布加权比较，另报告真实流量构成，避免把更容易的请求归功于模型升级．

14.10 将同设备全部日志放同一训练/验证/测试组，测试设备不能在训练中出现；若预测未来，还按时间留出．随机按行会让相邻传感器窗口、设备 ID 和同一次故障模式泄入两侧，结果可能依赖记忆而非跨设备泛化．应另测新设备和已知设备未来两个目标，不混称同一个泛化问题．

14.11 原 1000 任务中 300 经人工，700 自动含 70 错误．若 70 错误都被识别并转人工，则至少 370 个不同任务受干预，干预率 37%；700 不是最终无人工覆盖，真正完全自动且正确仅 630，即 63%．新增 70 的返工成功率与耗时题设未给，不能断言最终成功 970；若其成功率 r，则总成功 $900+70r$．成本应以 1000 全部提交任务平均，含自动尝试后返工双重成本．

14.12 保留任务类别、版本化输入引用、目标与实际行为、证据版本、工具动作/回执、终止原因及最小复现条件．敏感正文可放受控存储并用短期授权引用，日志保存哈希、字段类别或脱敏片段；凭据与无关个人信息不应记录．删除/保留期限、访问审计和回归样本用途需清楚，失败可复现不等于复制全部生产数据．

14.13 先核对文档版本、设备型号与运行条件，区分观测事实、相关线索、诊断假设和已确认原因．模型置信度可作为待评估信号，不能授予操作权限或证明物理安全．业务动作应由受控接口校验身份、对象、参数范围和前置状态，按风险执行审批，并由独立的确定性安全约束与联锁限制危险控制．记录证据来源、建议、批准主体、执行前状态、实际参数、执行结果及失败处置；对重复请求与状态变化规定幂等和一致性策略．执行后的独立状态核验用于确认效果，模型生成的成功描述不能代替设备证据．

第十五章　智能体架构

15.1 先预测 $\bar{b}_{t+1}(s')=\sum_{s} P(s' \mid s,a_t)b_t(s)$，再以似然更新

$$b_{t+1}(s')=\frac{P(o_{t+1} \mid s',a_t)\bar{b}_{t+1}(s')}{\sum_{u} P(o_{t+1} \mid u,a_t)\bar{b}_{t+1}(u)}.$$

分母为观测的边缘概率，保证归一化；零概率观测需模型外处理．若摘要删除了版本、来源或两个可能状态的区别，而这些会改变最优动作，它不是充分状态；语言流畅不能证明保留决策信息．

15.2 固定流程为解析文档 ID、授权检索、检查版本、组装证据、生成或拒答．受限 ReAct 允许看到版本冲突后查询历史版本，或首次检索不足后补检例外条款，并保留

步数、费用、期限和工具白名单. 只有这些新观测可能改变查询对象或终止决策时动态控制才有意义；盲目重复调用相同工具不是适应.

15.3 例如 A 确定适用合同版本，B 读取该版本价格，C 按价格计算总额，依赖 A 到 B 到 C. A 完成须有有效时间与来源，B 须匹配 A 修订，C 须金额/币种/数量验算. 若执行前发现合同版本变更，标记 B、C 依赖失效并重算；若任务问历史事实则继续固定历史快照. 不能只清空所有状态而重复已提交的业务动作.

15.4 单步工具可永不返回，所以有限步骤不限制时间；模型可无限无工具反思，所以有限工具次数不限制决策轮数. 控制器同时维护最大决策轮数、工具调用数、费用预留和绝对截止时间，并为每活动设置不超过剩余期限的超时. 等待审批释放执行资源，过期进入明确暂停/终止状态；循环不得通过子任务重置预算.

15.5 查询前最优值 0，查询后不扣成本的最优期望为 0.76；成本 0.8 时净信息价值为 $0.76-0.8=-0.04$，应不查询. 若查询期间资源变化，应把 $P(s' \mid s, \text{query})$ 纳入预测再决策，不能继续使用静态状态下的 $\dfrac{27}{29}$ 后验直接提交. 授权和资源版本仍要执行点校验.

15.6 保存命题、对象 ID、原始来源、来源修订、观察时间、适用时间区间、权限与核验状态. 晚收到的邮件可能转述旧配置，较早收到的正式发布可能已生效，因此观察时间不能代替有效时间或修订优先级. 冲突时按来源治理规则选择、回源核对或保留争议，不用向量相似度决定哪个事实有效.

15.7 恒定风险率给 $P(T>t+h \mid T>t)=1-\lambda h+o(h)$，故 $S'(t)=-\lambda S(t), S(0)=1$，解 $S(t)=e^{-\lambda t}$. 历史发生事实不因时间流逝而失效；按版本发布的配置通常跳变而非恒定随机失效. 库存风险也可随业务周期变化，需非恒定风险或事件驱动更新.

15.8 指定共同历史模型为 $n^2 m \sum_{r=1}^{R} r$，代入得 $16\times 100\times 6=9600$ 词元. 全互联三轮消息数为 $3\times 4\times 3=36$，与读入词元不是同一量. 差量消息保留事件 ID、版本和已确认游标，只发送新增信息，必要时获取检查点；漏游标或摘要丢失前提会导致错误决策，须能回放原证据.

15.9 若三者都复制同一错误数据库条目，其错误事件可以完全相同，多数表决不降低概率. 采用角色多样/相同与来源独立/共享的二乘二设计，固定任务、模型和总预算；报告条件错误相关、端到端成功及引用来源. 只有换角色而不换证据不能证明独立验证，独立来源也需排除共同转载祖先.

15.10 两者读版本 12；A 提交条件 version=12 成功并写 13，B 同条件失败，必须读 13 并重算或作冲突处理，不允许覆盖. 不同事件 ID 的只增日志可以集合合并，但余额、任务终态、批准摘要和同一字段替换需事务性约束. 对外执行还须隔离旧租约，CAS 只保护本地状态而非远端副作用.

15.11 发送后超时记未知，保留原动作键和摘要；查询业务端原键：已成功则补记回执，明确未提交且授权/期限仍有效则同键重试，查询不确定则继续等待或人工对账. 无法查询或不支持幂等时不可换新键盲重发. 补偿是另一个被授权动作，记录原预留与释放结果，并不抹去曾经预留的历史.

15.12 固定题集、模型、工具权限、总词元/费用/期限，比较单体与监督者配置；子体的成本计入总预算. 成功由外部目标状态判定，另报重复副作用、未知状态、迟到消息被采纳率、输入通信量和终止原因. 按题配对、按任务依赖/冲突切片；不能把更多预算的多体胜出归因于拓扑，监督者漏汇总也应作为系统失败计入.

15.13

不能. 观测信息、可达动作与预算同时变化，无法单独识别控制逻辑的作用. 应固定任务集合、模型版本、数据权限、工具版本、总预算和成功判据，仅改变所研究的控制机制，并采用同任务配对比较与重复运行. 若研究整体系统，则可保留不同配置，但须报告成本、权限、失败分母和人工介入；结论应限定为该配置下的端到端表现.

第十六章　智能体运行系统

16.1 知识回答返回论断、引用版本和支持状态；操作提案返回目标、规范参数、预期副作用与所需批准，不声称已完成；完成结果返回业务对象 ID、动作键、回执、状态与时间. 三者的证据分别是可定位来源、可校验计划、真实业务后置条件. 模型自报“完成”不能替代第三类证据.

16.2 令 $s_{t+1} = F(s_t, e_t)$ 且 F 确定，初态一致. 若重放到 t 时状态相同，则输入同一记录事件 e_t 后仍相同，由归纳所有状态一致. 若 F 读取未记录的时钟、网络或模型输出，第二次可能取得不同值，归纳步的相同输入前提失效. 应把这些结果记录为事件并回放，不重新调用以替代历史.

16.3 A、B 同时读版本 7；A 条件写 7 成功置 8，B 写 7 失败再读取 8，避免丢更新. 但 A 的旧租约可能在远端继续写，即使本地版本已经更新，数据库 CAS 也不能让远端自动停止. 工具端需验证递增隔离代次或稳定幂等键，并拒绝陈旧执行者；租约时间还需考虑时钟与延迟.

16.4 先登记已完成再执行，登记后崩溃会导致动作从未发生却被当完成；先执行再登记，执行后崩溃会在重试时重复. 若业务状态变更和去重结果可放在同一服务端事务，二者原子提交可避免两窗口. 跨服务事务不可得时，使用发件箱保证可投递、远端幂等保证重复提交不增副作用，并保留未知状态；本地记键本身不等于恰好一次.

16.5 发件箱未标发送会在恢复后再次投递，这是至少一次路径的正常行为. 接收端以逻辑动作键和摘要去重，返回既有结果；若删去去重，同一业务写可能执行两次.

发送标记提前写也可能丢消息，因此不能用“先标记”替代完整交付协议．去重保留窗口需覆盖重试和灾难恢复时间．

16.6 设总预算 B、已用 U、预留 R，接纳成本上界 b 必须满足 $U+R+b \leqslant B$；原子将 R 加 b 后并发接纳仍保持不变量．结算实际 c 且 $c \leqslant b$ 时，$U' = U + c, R' = R - b$，总和减少 $b - c \geqslant 0$．若 c 可能大于 b，原证明失效；需硬上限执行、分段追加预留或明确可承受超支政策，不能把事后记账称为预先预算保证．

16.7 状态可为已准备、已授权、发送中、未知、已确认成功、明确失败．工单已建但响应丢失时，以稳定动作键向工单系统查询并取得 ID 后转成功；查询明确未创建才允许同键重试．无查询能力时安排人工以审计字段核对，期间最终回答明确未知，禁止以新键创建第二张．人工核对本身也应留下证据与权限．

16.8 摘要覆盖主体、租户、工具版本、规范资源、参数、用途、数据级别、有效期及必要对象版本．批准签名绑定其哈希，任何参与字段变化都会产生新摘要，因此原批准失配．即使参数未变，所有权、余额或 ACL 版本改变也可能使动作不再授权，应在执行点重查或要求重新批准．摘要哈希不取代批准人身份验证．

16.9 误发通知可发送更正、撤销可撤销链接、停止后续推送并联系受影响接收者，均作为独立授权动作记录．已被读取或复制的内容无法保证收回，通知历史和隐私影响仍存在．不得把发送一条更正消息标为原事件从未发生；应分别报告业务缓解与信息披露残余．

16.10 轨迹回放验证给定历史输入下的控制器分支、预算和状态一致，不能证明外部接口现在可用．真实集成验证检查认证、协议及业务后置条件，但样本和受控环境有限．线上评估观察实际任务分布与依赖漂移，需要分层指标和故障归因，不能只靠 HTTP 成功率．三者互补而不能互相替代．

16.11 设两环节各以概率 0.1 失败，且由同一个错误目标解析触发，二者完全相关．各自成功率 0.9，独立乘积预测 0.81；真实端到端两者成功概率 0.9．反向也可构造互斥失败得到 0.8．通用式为 $P(E_1)P(E_2 \mid E_1)$，应测条件依赖而非假设层数越多就独立相乘．

16.12 保留任务状态模式版本、转换器版本、模型/提示/工具修订、原动作键、批准摘要及业务回执．旧任务可继续由旧解释器执行，或在检查点作可审计的状态迁移，迁移前确认无未决副作用；未知动作先对账．新模型只用于新决策，不能重写已记录决定．回滚应用版本不能撤销已经提交的工单或通知．

16.13

在轴对齐缩放且没有裁剪偏移的前提下，$x_v = \dfrac{1200 \times 960}{1920} = 600$，$y_v = \dfrac{600 \times 540}{1080} = 300$，故为逻辑坐标 $(600, 300)$．页面滚动后对象相对视口的位置可能变化，旧观测不再保证有效；应重新读取界面并定位．这里已经用视口与截图尺寸完成单位换算，不应再无依据地乘一次设备像素比．

16.14

由条件概率链式法则，$P = 0.98^{10} \approx 0.8171$. 该乘法使用条件概率，不需要另加独立假设. 回执丢失时动作可能已经生效；应保存目标对象与动作身份，查询业务结果或人工核实，再判断是否重试. 重新执行整条链可能重复创建对象，阶段检查点也不能替代写入幂等性.

16.15 Computer Use 是利用观察与输入操作软件的能力；CDP 是 Chromium 系浏览器的检测与控制协议；智能体工具协议负责向模型或运行系统描述并调用能力. 适配器可以把工具调用转为 CDP 命令，但兼容范围仍受底层驱动和版本限制，授权仍须由受信执行层按主体、对象及动作检查. 三者组合也不自动提供业务幂等与后置条件验证.

第十七章　优化及泛化

17.1 梯度为 $(2u, 8v)$，迭代因子为 $1-2\eta$、$1-8\eta$. 对任意初值趋零需两者模均小于 1，故 $0 < \eta < \dfrac{1}{4}$. 最小化最大模时令 $1-2\eta = -(1-8\eta)$，得 $\eta = \dfrac{1}{5}$，收缩因子 $\dfrac{3}{5}$. $\eta = 0$ 不移动；$\eta = \dfrac{1}{4}$ 的陡方向因子 -1，非零初值等幅振荡.

17.2 将正定矩阵正交对角化 $H = U\Lambda U^{\mathsf{T}}$，令 $a = U^{\mathsf{T}}\theta_0$. 则 $F(\theta_k) = \dfrac{1}{2}\sum_i \lambda_i (1-\eta\lambda_i)^{2k} a_i^2 \leqslant \rho^{2k}\dfrac{1}{2}\sum_i \lambda_i a_i^2$. 正定性保证各项权重为正且稳定区间内 $\rho < 1$；半正定时目标仍可能下降，但零特征方向不收缩，不能推出参数趋唯一零点.

17.3 独立有放回时 $\operatorname{Cov}(\bar g_B) = \dfrac{\Sigma}{B}$；批量变为 $4B$ 后任一固定投影方向的标准差减半，均方范数噪声减为四分之一. 若批内梯度完全相同，交叉协方差与对角项同量级，平均仍是同一个随机梯度，方差不随批量下降.

17.4 令误差递推 $e_{k+1} = (1-\eta h)e_k - \eta\xi_k$，条件零均值噪声方差为 $\dfrac{s^2}{B}$. 稳态二阶矩满足 $v = (1-\eta h)^2 v + \dfrac{\eta^2 s^2}{B}$，解得 $v = \dfrac{\eta s^2}{[Bh(2-\eta h)]}$；乘 $\dfrac{h}{2}$ 得损失 $\dfrac{\eta s^2}{[2B(2-\eta h)]}$. 需 $0 < \eta h < 2$. 减小步长降低噪声底，但收缩因子趋近 1，达到稳态更慢.

17.5 对称零均值预激活满足 $\mathbb{E}\phi(z)^2 = \dfrac{(1+0.1^2)\mathbb{E}z^2}{2}$. 下一层前向二阶矩乘子为 $\dfrac{256 s_W^2 (1.01)}{2}$；令其为 1 得 $s_W^2 = \dfrac{2}{256 \times 1.01} \approx 0.00773515$. 结论是对随机权重及输入取期望的尺度保持，不保证每次抽到的矩阵是等距映射.

17.6 取保留概率 q，半平方损失的额外项为 $\dfrac{1-q}{2q}\sum_j w_j^2 a_j^2$. 若掩码相关，读出方差

还含 $2\sum_{j<k}\dfrac{w_j w_k a_j a_k \operatorname{Cov}(m_j,m_k)}{q^2}$；半平方损失对应增加其一半．这里假设各掩码的边缘保留率均为 q，且条件于固定 a 求期望．

17.7 原目标 $F=L+\lambda\Omega$．整体改为 cF 时梯度乘 c，故用 $\eta'=\dfrac{\eta}{c}$ 可保留更新，目标内部 λ 不变．若接口只把数据损失改为 cL，需同时设 $\lambda'=c\lambda$，再用 $\dfrac{\eta}{c}$．若只改学习率而不改相对惩罚系数，优化的是另一目标；自适应优化器还需单独核查 ϵ 等状态，不直接套本结论．

17.8 逐记录随机划分泄漏客户特征，全部数据拟合标准化器又泄漏测试分布．可固定时间截点 $t_0<t_1$，训练仅用训练客户在 $t\leqslant t_0$ 的记录；验证用另一组未见客户在 $(t_0,t_1]$ 的目标；测试用再一组未见客户在 $t>t_1$ 的目标．任何历史特征只能取预测时可获得的数据；客户身份跨集合互斥，预处理仅在训练集拟合，超参数在验证后冻结．需报告新客户群体选择与上线群体是否相符．

17.9 待最小化项为 $(1-a)^2\mu^2+\dfrac{a^2\sigma^2}{n}$．求导得 $-2(1-a)\mu^2+\dfrac{2a\sigma^2}{n}=0$，所以 $a^*=\dfrac{\mu^2}{\mu^2+\frac{\sigma^2}{n}}$；二阶导非负，非退化时为正．若两项均为零任意 a 等价．μ、σ 未知，不能使用测试真值选 a；可在独立验证中选择或构造训练内估计并评估其误差．

17.10 先在相同评估模式下重算可比的数据损失，统一有效目标分母并区分正则项，核对训练与验证的预处理、标签和划分．再固定一个人工核验的小批次，关闭随机扰动，检查能否显著降低损失，同时记录梯度、更新幅度和非有限数值．小批次仍难拟合时，依次检查计算路径、学习率和参数更新，再通过受控容量变化检验表达能力；仅凭高损失不能认定容量不足．训练数据损失持续改善而独立验证退化，且已排除分布与统计口径变化，才支持过拟合判断．小批次拟合成功只能排除部分实现与优化故障，不能证明全量训练充分或模型能够泛化．

第十八章　训练数据工程基础

18.1 设 G 组独立、每组 $m=10$、各观测方差 σ^2，组内相关 $\rho=0.2$，总数 $N=10G$．均值方差为 $\dfrac{\sigma^2[1+(m-1)\rho]}{N}=\dfrac{2.8\sigma^2}{N}$，匹配独立同方差均值可定义 $N_{\text{eff}}=\dfrac{N}{2.8}$．它是特定均值估计量的方差等效，不刻画神经网络的函数复杂度、梯度相关结构和数据覆盖，不能直接解释为训练样本有效数量．

18.2 公共信封固定 dataset/record/group 身份、schema 版本、来源与许可、时间、内容哈希及权限．工具载荷保存有序消息、call ID、参数和结果引用，校验每结果匹配已有调用及角色次序；媒体载荷保存消息块和图片引用，校验资源存在、类型及 processor 版本；偏好载荷保存同一 prompt 下 chosen/rejected，校验共同条件、候选区分和分组原子性．未知类型进入隔离，不把三类字段拼成大量空列的万能表．

18.3 设接受事件 A 满足 $P(A \mid Z=z)=a(z)$，由条件概率 $P(Z=z \mid A)=\dfrac{a(z)P(z)}{P(A)}$，其中 $P(A)=\mathbb{E}_P a(Z)>0$. 两个领域原各 1000 条，接受率.9 和.1，则过滤后 900 与 100，份额.9/.1. 校正需要各领域准入后数量、接受机制和目标份额；可抽样或重要性加权，但被完全过滤的子群没有支持，不能仅靠无限权重恢复.

18.4 独立排列下 $\mathrm{Var}(\widehat{J})=\dfrac{J(1-J)}{m}=\dfrac{.25}{m}$，要求标准差不超过.025，故 $m \geqslant \dfrac{.25}{.025^2}=400$. 标准差是重复试验的波动尺度，不保证每次误差都在该范围内；确定概率容差还需另给置信水平并使用合适尾界或二项分布区间.

18.5 候选概率为 $1-(1-s^r)^b$. s=.6 时，(20,5) 得到 $1-(1-.6^5)^{20}\approx .8019$；(10,10) 得到 $1-(1-.6^{10})^{10}\approx .05885$. 相同 100 行签名，后者要求更长连续匹配，候选更少、精确复核量下降，却更易漏掉.6 相似记录. 候选量还取决于全库分布，不能仅靠本概率断言总运行成本.

18.6 可取 $A=\{1,2\}, B=\{1,2,3\}, C=\{2,3\}$，相邻 Jaccard 为 $\dfrac{2}{3}$，端点为 $\dfrac{1}{3}$. 阈值.6 下三个节点连通，分到同一评估组可避免任何高相似边跨集合；但 A 和 C 不达到阈值，不能因连通便把它们当成同一内容删除. 稳定代表去重需每个删除项与实际保留代表之间有直接依据.

18.7 冻结旧评估版本及其成员不做静默重排，先隔离引入连接的新记录和相关派生视图，调查新边是否揭示旧组已有泄漏. 若旧测试解释确受影响，将该版本标为受污染并停止用其作独立准入；重建组图、分组划分和全部下游版本，建立新测试基准. 历史分数保留但附失效说明；仅改新记录 split 不能自动修复已泄漏训练.

18.8 平方根计数归一化得到记录概率 p. 随后乘各来源平均长度，得到未归一化词元量 w，再归一化得到词元份额 q：

$$
\begin{aligned}
p &= \frac{(100,\ 10\sqrt{10},\ 10)}{110+10\sqrt{10}} \\
&\approx (0.70610,\ 0.22329,\ 0.07061), \\
w &= (5000,\ 2000\sqrt{10},\ 5000), \\
q &\approx (0.30629,\ 0.38742,\ 0.30629).
\end{aligned}
$$

若指定词元份额为 q_i，可设序列抽样概率 $p_i \propto \dfrac{q_i}{\mathbb{E}} L_i$，再按实际有效标签计数复核. 长度与过滤分布变化会使目标和物化结果偏离.

18.9 连续流 18 词元用 2 个 10 长块，填充 2，可按第一块 6 个 A 加 4 个 B、第二块 2 个 B 加 6 个 C 排列；B 被截断且允许跨样本读取会改变条件. 不可拆样本时任意两条长度 6 都装不进 10，需 3 块、填充 12. 独立注意力若也禁止拆样本，同为 3 块但还需块内样本 mask；若允许拆段可用 2 块并阻断跨样本读取，但第二段 B 是否保留第一段 B 上下文必须另定，不能声称仅靠 mask 保持原目标.

18.10 取块长 8：位置 [0,3) 来自数据集 D1 记录 A 源偏移 [5,8)，[3,4) 为插入 SEP，[4,6) 来自 D2 记录 B 偏移 [0,2)，[6,8) 为 PAD. 若 A 是提示、B 是回答，可设相应目标监督仅覆盖 B 及协议所需 EOS，具体按右移后目标而非输入编号判定. 每跨度存父记录和变换版本；查询块位置 5 落在 [4,6)，可反查 B 的偏移 1. SEP 记录生成规则、PAD 为空来源，不能伪称原文.

18.11 权重已包含某批更新，但游标仍指向该批之前，恢复后会再消费并更新一次；若反过来先推进游标再提交权重，则可能漏更新. 逻辑提交点应绑定权重、优化器/调度器、随机状态、分片游标、洗牌缓冲、未封存 packing 尾部及累计有效计数. 可先写不可变状态对象，最后原子切换包含全部摘要的提交清单，消费者只恢复完整同版本组.

18.12 血缘关系为 “源文档及别名 → 规范代表 → 训练视图 → 带 source span 的块 → 分片 → 训练任务 → 检查点”. 删除原对象后仍需阻止别名再次导入、定位代表的其他合法来源、撤销或重建受影响视图和分片，并标记使用过旧版本的任务. 未来消费撤销不自动去掉参数中已学影响；已有模型需按用途评估重训、遗忘或停用策略，保留受控审计而非继续公开敏感原文.

第十九章　训练稳定性

19.1 从 $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, v_0 = 0$ 展开得 $v_t = (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2$. 若各步 $\mathbb{E} g_i^2 = \mu_2$，则 $\mathbb{E} v_t = (1-\beta_2^t)\mu_2$，除以 $1-\beta_2^t$ 后无此初始化偏差. 无需时间独立性，但矩随时间变化时不等于当前真实矩.

19.2 约定矩和优化器内部步号一起重置，正 ϵ 保证零梯度坐标有定义，再取 $\epsilon \to 0^+$ 的纸面极限. 新步修正矩为 $(0,3)$ 与 $(0,9)$，方向 $(0,1)$，故

$$\begin{aligned}\theta_2' &= 0.999(0.899, -1.898) - 0.1(0,1)\\ &= (0.898101, -1.996102).\end{aligned}$$

原例为 $(0.83094, -1.94543)$. 若只清矩却保留步号 2，第二坐标方向为 $\dfrac{\sqrt{0.0199}}{0.19} \approx 0.74246$，得到 $(0.898101, -1.970348)$；这也是不同协议，不能含糊称为相同恢复.

19.3 最小化 $\dfrac{1}{2}\|u-g\|^2$ 且 $\|u\| \leqslant c$. 可行时 $u = g$；否则拉格朗日方程 $u - g + \lambda u = 0$ 给 $u = \dfrac{g}{1+\lambda}$，由边界得 $u = \dfrac{cg}{\|g\|}$. 逐坐标截断如 $(3,4) \mapsto (2,2)$ 改变比例 3:4；全局裁剪保持方向.

19.4 对正数 a，梯度序列变 ag_t 时，修正矩变 $a\widehat{m}_t, a^2\widehat{v}_t$，方向为 $\dfrac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \frac{\epsilon}{a}}$. 只有 $\epsilon = 0$

或稳定项可忽略时尺度近似不变；负缩放会翻转一阶矩方向．解耦权重衰减项本身不由这项缩放抵消．

19.5 两组正交参数坐标空间使全局范数为 5，阈值 2 给共同因子 $\frac{2}{5}$，两组范数变 1.2 与 1.6．分别裁到 2 后总范数为 $\sqrt{8} > 2$，且两组相对权重改变；它解决的是两个球约束，非同一个全局球约束．

19.6 若总更新数 $T \geqslant 2$，对 $t = 1, \ldots, T$ 取 $\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})[1 + \cos(\frac{\pi(t-1)}{T-1})]$，首步峰值、末步最低值．$T = 1$ 单独约定唯一一步为峰值或任务指定值，避免除零；只在成功参数更新后推进计数．

19.7 除模型、优化器、调度和随机状态外，保存当前未提交梯度、已累计有效目标数、微步位置、损失缩放器、数据消费与在途状态．还要明确梯度是和还是均值、是否已反缩放或同步．若不保存这些，可退回上一更新边界重放整个累积窗口，不能从中间游标继续而丢梯度．

19.8 以已消费游标为权威，预取不计训练进度．方案一取消在途预取并从已消费游标确定性重建三个批次；方案二保存预取内容及顺序、工作线程随机状态并恢复队列．不能既从已消费位置重读又恢复同一队列．随机增强无法重建时，应保存其种子或物化结果．

19.9 清单缺必要分片、大小/摘要校验不符或制品依赖不兼容时拒绝该版本，回退最近完整一致版本．输出明确失效类型、步骤和可能损失进度．允许迁移时先执行显式转换并产生新清单；不得跳过错误分片或混入其他步骤参数．

19.10 改变设备数会改变归约顺序、批次分组、数据分片和随机流，即使完整状态可重分片，浮点舍入也可不同．资产验收检查清单完整，语义验收检查有效目标和更新定义，数值验收比较容差内梯度/损失；逐位验收要求更强的确定性内核、顺序及随机配置，不能由前三级自动推出．

19.11 先定位非有限数值首次出现于输入、前向、反向还是优化器更新，并保留必要的批次与数值统计．若更新尚未提交且各设备一致确认异常，可跳过该次参数和优化器更新；动态损失缩放场景按协议调整缩放值．若参数或优化器状态已污染，应恢复一致检查点，包含参数、优化器、调度器、缩放器、随机状态与数据位置．降低学习率只可能缓解过大更新引起的不稳定，不能修复错误输入或已污染状态；梯度裁剪处理有限的大梯度，不能把 NaN 变为有效梯度．应分别记录成功更新数、跳步数与数据消耗，明确调度器按哪一计数推进，并在分布式训练中保持跳步与恢复决策一致．

第二十章　监督微调

20.1　$\log p(x,y)=\log p(x)+\log p(y\mid x)$，全序列 NLL 比回答条件 NLL 多出提示建模项．继续预训练或同时学习文档格式的生成任务可能有理由监督提示，但会分配梯度给提示分布并改变样本长度权重；不能把此目标称为纯回答 SFT．

20.2　明确序列为 (B,U,q,A,a,b,c,E,P,P)，共 10 位置，以下位置编号从 0 开始．$r=(0,0,0,0,1,1,1,1,0,0)$，九个预测位置的 $m=(0,0,0,1,1,1,1,0,0)$．有效配对为位置 3 预测 a、4 预测 b、5 预测 c、6 预测 E；填充不作为目标，真实终止 E 必须保留．

20.3　由结构化消息跨度和实际有效长度判定角色，而不是按 token ID 过滤．真实消息结束位置即便编号等于 PAD，也保留 $r=1$；长度之外填充 $r=0$．再以 $m_t=r_{t+1}$ 构造监督，并结合注意力可见性，避免删除全部同编号词元．

20.4　两条 NLL 和为 8 和 12．词元等权为 $\dfrac{20}{16}=1.25$，样本等权均值为 $\dfrac{2+1}{2}=1.5$，平均序列 NLL 为 $\dfrac{8+12}{2}=10$（若不平均则总 NLL20）．前者样本权重 $\dfrac{1}{4}$ 与 $\dfrac{3}{4}$，中者各 $\dfrac{1}{2}$ 且每词元权重分别 $\dfrac{1}{8}$ 与 $\dfrac{1}{24}$；后者每词元权重均 $\dfrac{1}{2}$ 但总尺度更大．

20.5　三个累积前缀分别包含助手轮 1、轮 1 至 2、轮 1 至 3，若每个前缀监督全部助手，轮 1 出现 3 次、轮 2 出现 2 次、轮 3 出现 1 次．方案一仅用完整三轮序列并监督三轮；方案二用三个前缀但各只监督新增末轮．若按样本均值再归约，仍需核对各轮有效长度造成的权重差异．

20.6　回答位置隐状态通过因果注意力依赖提示嵌入，链式法则给出

$$\frac{\partial L}{\partial e_j}=\sum_{t\in\mathcal{S}}\frac{\partial L}{\partial h_t}\frac{\partial h_t}{\partial e_j}.$$

j 自身没有直接标签只删除局部输出损失，不令右端为零；除非嵌入冻结、梯度截断或计算依赖本身为零．

20.7　正确保留质量 $0.7\times0.8=0.56$，错误保留质量 $0.3\times0.1=0.03$，保留率 0.59，纯度 $\dfrac{0.56}{0.59}=\dfrac{56}{59}\approx94.92\%$．只要 r/f 是目标总体上的真实条件率，公式无需独立性；但相关错误会使跨题/领域外推和置信区间失效，且高纯度不消除集中系统错误．

20.8　记录消息角色、事件顺序、已接受事实和来源．用户修正后检查后文采用新值并标记旧值被替代；信息不足时检查是否提出必要澄清而非编造；工具返回需绑定请求、状态和内容，并检查助手没有将工具失败解释为成功．以独立题族划分训练与评估，矛盾和缺失可标为不可判定，不强造唯一金标．

20.9　用户问题可以引用与回答相同的句子，助手也可能多次重复它；字符串首次匹配会选中提示．模板还可能插入控制标记，子词边界也不等于字符边界．应从结构化

消息和生成区间记录目标跨度，映射到词元索引，并验证模板渲染一致及停止标记归属.

20.10 旧任务梯度 $g_o=(1,0)$. 新任务若 $g_n=(2,0)$，一步更新 $-\eta g_n$ 使旧损失一阶变化 -2η；若 $g_n=(-2,0)$ 则变化 $+2\eta$. 正内积支持局部改善，负内积提示干扰；非线性、高阶项和后续更新可改变趋势，不能仅用一次梯度角度证明长期保持.

20.11 条件似然可在独立参考答案集按明确权重计算 NLL，但它仍依赖参考写法与词元长度；生成质量用可执行验证或盲评量表覆盖多种合法答案，并报告失败/拒答. 模型可能给某一金标较低概率却生成另一正确答案，也可能金标似然高但自由采样时偏离，因此须同时报告.

20.12 先冻结新领域、通用能力和风险切片及可接受退化阈值；只在验证集筛选同时满足保持约束的候选，再比较领域收益与成本. 正式测试仅在选择及阈值固定后使用；若根据测试结果反复选模型，它转为开发集，需另建独立测试. 高风险失效独立设门槛，不由领域均分抵消.

第二十一章　参数高效微调

21.1 参数量 $r(d_i+d_o)=16(4096+1024)=81920$；基座参数 $4096\times1024=4194304$，比例 $\dfrac{5}{256}\approx1.9531\%$. $2rd$ 仅适用于输入输出同宽，不能随意取 4096 代入.

21.2 $dY=s(dB)AX+sB(dA)X+(W_0+sBA)dX$，代入 $dL=\operatorname{tr}(G^{\mathsf T}dY)$ 并循环移位，得到 $\nabla_B L=sGX^{\mathsf T}A^{\mathsf T}$、$\nabla_A L=sB^{\mathsf T}GX^{\mathsf T}$、$\nabla_X L=(W_0+sBA)^{\mathsf T}G$. 其形状分别 $d_o\times r, r\times d_i, d_i\times n$；冻结 W_0 不删除最后一式.

21.3 第一次后 $A_1=(1,-1), B_1=(0.1,0.1)^{\mathsf T}$，$y-t=(0.9,0.9)$，$A_1x=-1$. 故 $\nabla_B=(-0.9,-0.9)^{\mathsf T}$，$\nabla_A=(0.18,0.36)$. 步长 0.1 得 $B_2=(0.19,0.19)^{\mathsf T}$，$A_2=(0.982,-1.036)$. $A_2x=-1.09$，输出 $(0.7929,1.7929)$，损失 $0.7929^2=0.62869041$. 若先改 B 再计算 A 梯度，则不再是同一点的一次普通梯度下降.

21.4 由结合律 $(BR)(R^{-1}A)=B(RR^{-1})A=BA$. R 可旋转或缩放通道，故单通道不具唯一语义，乘积才不变. 不同适配器因子处于不同基时直接平均没有不变意义；应先对完整增量求和或对齐后再近似.

21.5 $\Delta W=BA$ 的每一行是 A 行的线性组合，故行空间包含于固定的 $\operatorname{row}(A)$；若输入 x 位于 A 零空间，则增量对 x 永远为零. 共同训练 A/B 允许改变该子空间，但仍受秩上界 r 限制，优化非凸并不保证找到任意期望子空间.

21.6 B 零初始化时 $\nabla_A=0$，一步增量为 $-\eta s^2GX^{\mathsf T}A^{\mathsf T}A$. 若 A 各元素独立零均值、方差 σ^2，则 $\mathbb{E}A^{\mathsf T}A=r\sigma^2I$. 固定 $s=\dfrac{\alpha}{r}$ 且固定 alpha/方差，期望幅度正比 $\dfrac{\alpha^2\sigma^2}{r}$；固定 s 则正比 $s^2r\sigma^2$. 若初始化方差也随 r 变化，必须把它代入，不能只按 r 判断.

21.7 取 $B_1 = e_1, A_1 = e_1^\top$ 与 $B_2 = e_2, A_2 = e_2^\top$，增量和为 2 维单位阵、秩 2. 但 $(B_1 + B_2)(A_1 + A_2)$ 是全一矩阵、秩 1，并多出 $e_1e_2^\top + e_2e_1^\top$. 拼接 $[B_1, B_2]$ 和纵向堆叠 A 可精确表示增量和.

21.8 码值固定时只把 $D(W_q)$ 看作常矩阵，无需对离散码求导；但 $\nabla_X L = [D(W_q) + sBA]^\top G$ 仍包含基座项. 对整条基座分支使用无梯度执行会丢失该项，影响更早层的适配器或可训练输入，从而改变目标梯度.

21.9 输入软提示为 $pd = 20 \times 2048 = 40960$. 若每层键和值各有 d 维，直接前缀为 $2Lpd = 1966080$，是前者 48 倍. 另计占用的上下文/KV、提示编码器（若有）、注入位置、推理内核支持、加载切换和任务质量；不能只以参数量选方法.

21.10 若 $y = x + U\phi(Dx + b_d) + b_u$ 且 ϕ 为恒等，则 $y = (I + UD)x + (Ub_d + b_u)$，是仿射映射；只有附加偏置 $Ub_d + b_u = 0$ 时才是纯线性增量. 保留 ReLU 等非线性时雅可比随 x 变化，一般无法由单一常矩阵和偏置表示. 固定偏置不等于偏置为零，原题已据此澄清.

21.11 取 q=1、键 $(0, 1)$、值 $(0, 1)$，原输出为 $\dfrac{e}{1+e}$. 把键乘 2 后为 $\dfrac{e^2}{1+e^2}$；把原输出乘 2 则为 $\dfrac{2e}{1+e}$，不仅不同，后者还超过值凸包. 键缩放改变概率分配，不能移到 Softmax 后的值缩放.

21.12 稳定信息如职责和固定风格进入版本化训练档案；租户、权限与部署版本由服务配置提供；日期和工具结果来自带来源的实时观察. 按提示家族划分数据，加入普通数学/翻译等非触发对照；测试身份、能力、动态事实和越权分别计分. 出现旧角色泄漏、过度自我介绍、伪造工具成功或通用能力越界退化时回退完整适配器/模板组合，授权始终由执行器裁决.

21.13 矩阵乘积满足 $\operatorname{rank}(BA) \leqslant \min(\operatorname{rank}(B), \operatorname{rank}(A)) \leqslant r$. 取 $W = I_d$、$r = 1$，$B = A^\top = e_1$，则 $W + BA = \operatorname{diag}(2, 1, \ldots, 1)$ 仍为满秩. 低秩约束作用于特定权重的更新方向，多层非线性和多个适配位置共同决定最终函数，不能由单个增量秩直接推断整体任务能力. 增大 r 扩展可更新方向并增加参数与优化成本，但验证效果还受数据、目标层、正则化和优化充分程度影响，应在相同评估与资源约束下比较.

第二十二章　训练工程优化

22.1 冻结基座占 1.2×10^9 字节，可训练状态占 3.2×10^8 字节，总量为 5.52×10^9 字节. 将可训练参数减半只节省 1.6×10^8 字节；激活、临时空间及运行时仍需要单独测量.

22.2 正确梯度为 $\dfrac{3 \times 2 + 9 \times 6}{12} = 5$，微批均值平均为 4. 累积步数只计微批次，不能表示每个微批次的有效监督数量；回答掩码和变长样本会使这些数量不同.

22.3 取 $k=\sqrt{100}=10$，近似激活存储为 $20(10+10)=400$ MiB，全部保存为 2000 MiB. 峰值还需加入参数、梯度、优化器状态、临时张量、非均匀层开销和运行时工作空间；同时评估额外前向计算.

22.4 先核对可训练参数集合、基座精度和优化器状态，在目标长度下分别记录前向、反向与更新阶段的峰值，覆盖首次状态分配和稳定更新步. 冻结权重仍可能参与输入梯度传播，应检查实际反向路径. 分别改变长度、微批量和重计算配置，判断增长是否来自激活；基座及复制状态占主导时，再评估相应权重或状态分片. 适配器优化器状态较小时，仅分片该部分收益有限. 普通数据并行通常保留完整副本，不能直接消除单卡容量不足. 比较方案时保持有效监督预算与质量要求，复核梯度累积的归一化、全局批量、通信和恢复语义，并同时记录峰值显存、墙钟时间与设备时.

第二十三章　预训练原理

23.1 平均 NLL 为 $\dfrac{\log 2+\log 4+\log 8}{3}=\log 4$，困惑度 4. 正确目标概率的几何平均为 $(\dfrac{1}{64})^{\frac{1}{3}}=\dfrac{1}{4}$，PPL 是其倒数；不是概率算术平均的倒数.

23.2 令两微批平均梯度为 g_1,g_2，正确更新方向为 $\dfrac{g_1}{4}+\dfrac{3g_2}{4}$. 等权错误方案给短批权重 $\dfrac{1}{2}$ 而非 $\dfrac{1}{4}$，故其单词元贡献放大 2 倍，长批缩为正确值的 $\dfrac{2}{3}$.

23.3 将各文档记为 (B,a,b,c,E) 与 (B,d,e,f,E). 有效配对分别为 $B\to a, a\to b, b\to c, c\to E$ 和 $B\to d, d\to e, e\to f, f\to E$，共 8 项；不预测跨文档的 $E\to B$. 注意力为按文档分块的下三角，位置编码重置与否须跟训练协议一致；EOS 本身不能自动实现隔离.

23.4 等概率抽文档时贡献为 $0.5\times 256:0.5\times 1024=1:4$，词元比例 20% 与 80%. 要词元等权，解 $256p=1024(1-p)$ 得 $p=0.8$；该式以平均有效目标长度稳定为前提.

23.5 对象未出现概率为 $(1-\dfrac{1}{n})^{2n}\to e^{-2}$，故期望不同对象比例为 $1-e^{-2}\approx 86.47\%$. 两次无放回完整遍历覆盖 100%，有放回的 2n 次消费仍会重复和遗漏；消费次数不是信息覆盖量.

23.6 $N_*\propto C^{\frac{\beta}{\alpha+\beta}}=C^{\frac{3}{5}}$，$D_*\propto C^{\frac{2}{5}}$；损失余量指数为 $-\dfrac{\alpha\beta}{\alpha+\beta}=-\dfrac{1}{5}$. 这些是固定幂律、系数和 $C=\kappa ND$ 前提下的连续最优关系，不用于数据重复或训练效率随规模改变而未重新拟合的情形.

23.7 拉格朗日条件给 $\alpha AN^{-\alpha}=\beta BD^{-\beta}$. 平衡的是每对数预算变化的边际贡献，故两余量之比为 $\dfrac{\beta}{\alpha}$；只有指数相等时才相等，不能从“最优”直接推出两损失项相同.

23.8 预算给 $nd = 144$，边际条件仍为 $d = 16n$，故无约束最优 $n = 3, d = 48, L = 2 + \frac{2}{\sqrt{3}} \approx 3.1547$. 若 $N \leqslant 2 \times 10^9$，则 $n \leqslant 2$，最优在边界 $n = 2, d = 72$，$L = 2 + \frac{1}{\sqrt{2}} + \frac{4}{\sqrt{72}} \approx 3.1785$；假定数据可用且允许此训练长度.

23.9 例如只恢复 15% 被遮蔽位置的双向模型与预测所有后继位置的因果模型，条件信息、计分位置和目标难度均不同. 前者可看右文，且处理量与监督量不同. 应固定下游任务、数据预算和计算口径比较，而不是把较低的条件 NLL 当作更强模型.

23.10 可选 $N \in \{0.5, 1, 2\}$ 与 $D \in \{4, 16, 64\}$ 的笛卡尔网格并独立训练，保留相同分布和适合各规模的优化条件，观察固定 D 时 N 效应与固定 N 时 D 效应. 一个运行的检查点共享参数规模、数据前缀和优化历史，提供学习曲线却不能识别另一个参数维度；需记录拟合残差和外推区间.

23.11 参数上限限制驻留和每次推理成本，可以选择较小模型并延长训练，即使偏离训练计算最优. 令总成本为训练成本加预期请求量乘推理成本，在质量约束下求可行最小值；请求量、缓存和延迟假设必须给出，不存在不依赖部署负载的统一最优规模.

23.12 同一权重若重启数据游标，会重复或遗漏样本；混合时钟重置会改变后续分布；监督计数丢失会改变累积归约与预算. 恢复应同时绑定权重、优化器/调度器、游标/随机状态、混合日程及已成功更新计数，并在更新边界提交；仅加载权重只证明资产可读.

23.13 先固定独立领域验证集与通用能力评估，按数据源统计去重后规模、累计训练词元和重复次数. 以可比计算预算安排重复训练、不同混合比例与模型规模的对照，另保留较早检查点作为停止方案；报告计算量和实际时间，避免把重复词元视作新增独立信息. 比较领域收益、通用能力退化及各数据切片的不确定性. 额外轮数仍带来稳定独立验证收益且满足通用能力约束时，可继续；验证收益停滞或退化时，应优先检验数据覆盖与目标匹配，而不能只按训练损失下降追加预算. 多轮训练可能有效，但其边际收益需要测量.

第二十四章　强化学习基础

24.1 环境状态可包含搜索索引版本、网页实际内容、访问权限及已执行工具副作用；观测是返回片段、状态码和已读文本；动作是搜索查询、读取目标或最终回答. 未返回网页、更新时刻和后端负载对策略不可见，因此历史 h 或信念状态承载条件信息，不能把局部搜索摘要当成全局真实状态.

24.2 第一步奖励 0，第二步奖励如原例，故 $J = \gamma[4pq + 2(1-p)] = 2.25$. $V(s_L) = 3, V(s_R) = 2$ 仍按各状态自身时钟计；$\partial_\theta J = 0.9 \times 0.25 = 0.225$，$\partial_\psi J = 0.9 \times 0.375 =$

0.3375. 初始分布目标的策略梯度在 t=1 项必须带外部 γ^1，不能只在局部回报中折扣后遗漏它.

24.3 由 $G_t = r_t + \gamma G_{t+1}$ 条件取期望，得 $V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^\pi(s')]$. 两价值函数经同一 Bellman 算子后的无穷范数差至多 $\gamma\|V - W\|_\infty$. $\gamma = 1$ 时仅非扩张，不是严格压缩；有界有限回合仍可倒推，但无限过程需另有终止或收敛条件.

24.4 三条轨迹 LS、LF、R 的概率为 $\frac{3}{8}$、$\frac{1}{8}$、$\frac{1}{2}$. 对 psi 得分分别 $\frac{1}{4}$、$\frac{-3}{4}$、0；乘回报 4、0、2 再取期望为 $\frac{\frac{3}{8} \times 4 \times 1}{4} = \frac{3}{8}$，与 $4pq(1-q)$ 一致. 未访问 sL 的轨迹无 psi 动作项，不能赋予它虚构的第二步得分.

24.5 固定状态时 $\sum_a \pi(a|s) b(s) \nabla \log \pi(a|s) = b(s) \nabla \sum_a \pi(a|s) = 0$，需阻断基线梯度. 反例：二动作奖励 1/0，取动作基线等于奖励，则估计处处为 0，而原目标 $J = p$ 的 logit 梯度为 $p(1-p) > 0$. 动作相关项不自动消失.

24.6 令得分 g_i 均值为零，组内独立同分布. $\mathbb{E}[(r_i - \bar{r})g_i] = \mathbb{E}[r_i g_i] - K^{-1}\mathbb{E}[r_i g_i]$，其余交叉项由独立性和得分零均值为零，故比例 $1 - \frac{1}{K}$. K=1 时优势恒 0，完全无奖励梯度；留一基线在 K≥2 且独立时可避免这项偏差.

24.7 原例残差为 0.4、1.2，$\gamma = 1$. lambda=0 时 $(\widehat{A}_0, \widehat{A}_1) = (0.4, 1.2)$；lambda=1 时为 $(1.6, 1.2)$. 前者仅用一步 TD、依赖价值估计，后者此终止轨迹等于完整回报减价值. 二者不是精确平均优势，仍含轨迹采样波动.

24.8 若环境未终止，应在最后残差保留 $\gamma V(s_{t+1})$ 自举；但优势递推不得越过缓冲边界连接到另一条轨迹，应在该片末设后续未知优势为 0 或使用明确定义的延续估计. 真实终止价值为 0，切片终点价值一般不为 0.

24.9 权重 $1.1^{10} \approx 2.593742$. 更长序列的乘积可能迅速变大或变小，少数轨迹主导估计；有限均值并不保证低方差. 可监控有效样本量和权重尾部，限制样本陈旧性；截断权重降低方差但引入偏差，应明确为近似目标.

24.10 取 epsilon=0.2. 正优势 +1 时 r=1.3 用常数 1.2、r=0.7 仍用 0.7；负优势-1 时 r=0.7 用-0.8、r=1.3 仍用-1.3. 裁剪只抑制某些有利方向的奖励增长，既不把参数投影到概率比区间，也不阻止共享参数或其他损失使比值继续越界.

24.11 固定终点 R>0，原始序列期望按每回答一次 R 计，长度本身不改奖励；R/T 偏好更短；终点折扣为 $\gamma^{T-1}R$（以动作 0 起计），gamma<1 也偏短；每步成本 c 则回报 $R - cT$，可在收益值得时保留更长过程. 若 R<0，除以 T 可能反而让长序列负值更接近零，方向不能照搬正奖励.

24.12 重要性分母必须是实际生成该样本的行为分布 mu，才能满足 $\mathbb{E}_\mu[(\frac{\pi}{\mu})f] = \mathbb{E}_\pi f$；参考模型只是正则锚点，通常不等于 mu. 硬截断令某些 mu 为 0 时无法用这些样本覆盖 pi 的正质量，支持条件失败；应匹配实际采样分布或明确采用有偏代理.

第二十五章　人类反馈学习

25.1 令 $d = r_w - r_l$，$L = -\log\sigma(d)$，则 $\dfrac{\partial L}{\partial d} = \sigma(d) - 1$. 对优选为 $\sigma(d) - 1$，对拒选为 $1 - \sigma(d)$，相加为零. 目标只识别差值；两输出共同平移不会改变该样本损失.

25.2 四回答只比较 a 与 b、c 与 d. 分别对第一连通分量加常数 u、第二分量加 v，所有已观测差值不变，故 u-v 不可辨识. 没有跨分量比较不能由数据确定 a 与 c 的相对奖励；正则可能选出数值解，但那是额外约束而非数据证据.

25.3 在参考正支持上令 $\pi^*(y) = \dfrac{\pi_0(y)e^{\frac{r(y)}{\beta}}}{Z}$. 利用恒等式

$$\log(\frac{\pi}{\pi_0}) = \log(\frac{\pi}{\pi^*}) + \frac{r}{\beta} - \log Z,$$

代入可得

$$J(\pi) = \beta\log Z - \beta\mathrm{KL}(\pi\|\pi^*).$$

由 KL 非负性得到最优策略. 有限支持及有限奖励确保 Z 有限；参考为零处不能随意赋正质量，否则 KL 无穷.

25.4 未归一化质量为 $(0.8, 0.2e^{\frac{\log 4}{2}}) = (0.8, 0.4)$，Z=1.2，策略 $(\frac{2}{3}, \frac{1}{3})$. 期望奖励 $\dfrac{\log 4}{3} \approx 0.462098$，KL 为 $\dfrac{2}{3}\log(\dfrac{5}{6}) + \dfrac{1}{3}\log(\dfrac{5}{3}) \approx 0.048728$，目标约 0.364643，等于 $2\log 1.2$.

25.5 同提示加 c 使各未归一化质量同乘 $e^{\frac{c}{\beta}}$，归一化后策略不变，J 增加 c. 不同提示各有不可辨识偏移时，跨提示直接比较奖励均值会受任意常数影响；须有共同校准或只作同提示比较，不能把排序奖励当绝对效用.

25.6 设相对间隔 Delta，损失 $-\log\sigma(\beta\Delta)$，对参数梯度为 $-\beta\sigma(-\beta\Delta)\nabla\Delta$. Delta=0 时系数-beta/2，很正时趋 0，很负时趋-beta. 链式梯度提高相对对数概率差，但共享归一化使各回答绝对概率变化不能仅由符号判断.

25.7 参考固定，策略从 $(p_w, p_l, p_o) = (0.4, 0.3, 0.3)$ 改为 $(0.35, 0.15, 0.5)$. 优选概率下降，但比值从 $\dfrac{4}{3}$ 升至 $\dfrac{7}{3}$，相对间隔增 $\log(\dfrac{7}{4}) > 0$，DPO 损失下降. 第三个回答吸收质量，说明排序改善不保证优选绝对概率增加.

25.8 序列概率为条件概率乘积，其 log 为逐词元 log 之和. 平均后得到 $\frac{\log p(y)}{T(y)}$，指数为 $p(y)^{\frac{1}{T(y)}}$，对所有序列求和一般不为 1；若重新归一化还需新的配分函数. 因此长度平均目标属于新的偏好尺度，不可直接代回原最优策略恒等式.

25.9 对同一初始提示，$\mathrm{KL}(\pi\|\pi_0) = \mathbb{E}_{y\sim\pi}\sum_t \log[\frac{\pi(y_t|h_t)}{\pi_0(y_t|h_t)}]$. 再按 pi 的历史分布条件取期望，得 $\sum_t \mathbb{E}_{h_t\sim\pi}\mathrm{KL}(\pi(\cdot|h_t)\|\pi_0(\cdot|h_t))$. 若轨迹从 mu 来且不校正，外层历史分布与动作分布都变，通常不再等于此序列 KL.

25.10 对 G 个非全等样本，总体方差分母 G，样本方差分母 $G-1$；后者标准差乘 $\sqrt{\frac{G}{G-1}}$，在 $\epsilon = 0$ 时优势幅度乘 $\sqrt{\frac{G-1}{G}}$. 若分母保留正 ϵ，比例应为 $\frac{s_{\text{pop}}+\epsilon}{s_{\text{sample}}+\epsilon}$. 均值仍零，但训练尺度不同；全等时需明确 epsilon 或跳过规则，不能除零.

25.11 保存参考权重/适配器摘要、分词器、完整模板、精度、截断/停止协议、提示与回答词元、计分掩码、位置/可见性及是否求和. 模板或 EOT 变化会改变词元/条件/计分区域，原概率即失效；若只改变不参与前向的报告名称，缓存可保留. 不能只用原始问题字符串作键.

25.12 固定原问题和候选，制作保持事实相同的长短改写测标注长度偏好；对评分器加入事实错误但风格合规的对抗候选测漏洞；冻结问题及采样预算重评旧新策略，隔离候选分布漂移. 由独立事实核验者盲评并分长度/领域报告. 若改变验证器，保留版本交叉矩阵，不能仅凭奖励曲线诊断原因.

第二十六章　可验证推理训练

26.1 算术符号等价只证明在声明定义域和假设下式子相同，不能证明建模符合题意；代码单测通过只覆盖给定输入/环境；数据库工具成功只说明语句执行或事务提交，仍要验证目标行、授权范围和最终业务不变量. 三者都需区分执行器异常与候选失败.

26.2 筛后纯度为 $\frac{pr}{[pr+(1-p)f]}$. 要求至少 0.9 得 $f \leqslant \frac{pr(1-0.9)}{0.9(1-p)} = \frac{0.0019}{0.882} \approx 0.0021542$，约 0.2154%. 低基率下即便正确通过率很高，少量误收也会主导保留集；条件率必须在目标难度分布校准.

26.3 均值为 $\frac{1}{4}$，平方偏差和为 $\frac{3}{4}$. 总体标准差为 $\frac{\sqrt{3}}{4}$，样本标准差为 $\frac{1}{2}$，三种优势

分别为

$$A_{\text{pop}} = (-\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \sqrt{3}),$$
$$A_{\text{sample}} = (-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \frac{3}{2}),$$
$$A_{\text{LOO}} = (-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 1).$$

前两者采用不同的随机尺度归一化；第三者使用留一基线且未标准化，不应互称同一优势.

26.4 独立二元奖励下概率为 $1 - p^G - (1-p)^G$. p=.05 时 G=2 为.095；G=8 约.33657957；G=32 约.80628852. 可比较每单位生成词元的非零方差组数及最终质量，而非只最大化混合组概率；大 G 生成代价近线性增加，奖励相关时该式失效.

26.5 令 $g_i = \nabla \log \pi(y_i|x)$，条件于 x 时均值零. 展开 $\mathbb{E}[(r_i - G^{-1}\sum_j r_j)g_i]$，自身项留下 $(1 - \frac{1}{G})\mathbb{E}[r_i g_i]$，j 不等 i 项需候选条件独立才消失. 共享搜索树导致奖励与另一候选得分相关，不能直接沿用；需按联合采样分布重新推导或用独立对照.

26.6 原序列比为 $1.5 \times 0.5 = 0.75$，平均词元比为 1. 另一回答取 $(1, 1)$，平均仍 1 但序列比 1. 乘积与平均捕获不同对象，序列长时差异会累积.

26.7 按回答平均、beta=0，四条平均项为 1、1.1、-1、-.9，故 J=.05，损失-.05. 约定在饱和边界对比值取平坦分支导数 0：第一条局部 log 概率导数 $(0, \frac{.9}{8}) = (0, .1125)$；第二条比 1.1 都在上边界，导数全 0；第三条为 $(0, -\frac{1.1}{8}) = (0, -.1375)$；第四条比.9 处于负优势下边界，导数全 0. 最小化损失再反号. 边界可取其他合法次梯度，实现需明示；J 数值不因此改变.

26.8 令 $x = \frac{\pi_{ref}(a|h)}{\pi(a|h)} > 0$，由 $\log x \leqslant x - 1$ 得 $\kappa = x - 1 - \log x \geqslant 0$. 在当前 pi 完整支持上采样，$\mathbb{E}_\pi(x-1) = 0$，故期望为 $\mathrm{KL}(\pi \| \pi_{ref})$. 行为 mu 样本若不做校正，此归一恒等式不成立；共享支持或绝对连续性不足时须另外处理零概率.

26.9 按回答平均时同一负优势在长度 2 中的每位置系数为 $\frac{A}{2G}$，长度 20 为 $\frac{A}{20G}$，长回答每位置惩罚缩小 10 倍；按批量词元平均为 $\frac{A}{\sum_i T_i}$，两者每位置一致；固定分母为 $\frac{A}{GT_0}$，也一致但总体尺度不同. 还要乘实际概率比及裁剪分支；这些局部系数提示长度动力学，不能单独证明最终长度必然增长.

26.10 $k = 1$ 为 $\frac{4}{20}$=.2；k=2 为 $1 - \frac{\binom{16}{2}}{\binom{20}{2}} = \frac{7}{19} \approx .368421$；k=5 为 $1 - \frac{4368}{15504} = \frac{232}{323} \approx .718266$. 它估计候选中至少一条成功，产品多数投票仍需正确答案获得最多票；错误答案也可能聚集，故不能当投票返回准确率.

26.11 以问题为单位记录全部尝试顺序、候选、验证结果、实际消耗、停止原因、最终返回及独立真值. 直接统计策略返回成功率、成本分布和超时率. 样本数取决于先前成功, 固定 n 的组合无偏性前提已改变; 可以直接评估服务策略, 不能把该日志当作预先固定 n 的独立样本池.

26.12 任务要求 “仅取消指定未发货订单”, 工具却成功取消了另一订单, 或只成功查询而未取消. 终点验证需检查目标订单 ID、原状态、授权主体、事务提交及指定状态变化, 并确认无额外行被改. 奖励所有成功调用会鼓励无效操作, 应以受控的目标状态和副作用约束给奖励.

26.13 先按原题来源/模板族划分, 再生成教师轨迹; 训练生成不能访问封存测试答案. 验证器阈值在独立开发集确定, 保留版本; 教师检索和知识库也需查污染. 测试只用于冻结方案, 若按测试失败继续合成数据或改停止策略, 应另设新测试集; 字符串去重不足以覆盖改写与同模板题.

26.14 固定问题分布重评新旧策略; 固定候选做长度保持/事实保持改写, 测评分器对冗长的响应; 用独立执行或人工真值核对高奖励失败例. 报告长度、分项奖励、每题难度、独立正确率和生成预算. 仅问题混合改变时用旧混合重加权作对照; 评分漏洞需更新验证器并重新验收, 不以更高平均奖励作解释.

第二十七章　模型规模及资源预算

27.1 设查询头总宽为 $h_q d_h$, 键值总宽为 $h_{kv} d_h$, 输入及输出均为 d, 无偏置. 四个矩阵参数之和为 $d h_q d_h + 2 d h_{kv} d_h + (h_q d_h) d = 2 d d_h (h_q + h_{kv})$. 只有 $h_q d_h = d$ 时首尾两项才合为 $2d^2$; 若键和值头宽不同, 还应把两项分别写成 $d h_{kv} d_k$ 和 $d h_{kv} d_v$, 输出矩阵宽度相应修改.

27.2 保留 12 层、每层两次归一化及末端一次归一化; 补充约定末端也改为带缩放和偏置的 LayerNorm, 投影仍无偏置. 矩阵主干为 135266304, 两个嵌入/输出矩阵为 65536000, 25 次归一化各含 $2d = 2048$ 个参数, 合计 200853504. 若仅按题干字面修改层内 24 次而末端仍为 RMSNorm, 则总量为 200852480; 两答案相差末端新增的 1024 偏置, 必须声明此配置.

27.3 令 $Y = WX$, $W \in \mathbb{R}^{d_o \times d_i}$、$X \in \mathbb{R}^{d_i \times n}$、上游 $G \in \mathbb{R}^{d_o \times n}$. 前向约 $2 d_o d_i n$ FLOP; $\nabla_W L = G X^{\mathsf{T}}$ 形状 $d_o \times d_i$, $\nabla_X L = W^{\mathsf{T}} G$ 形状 $d_i \times n$, 二者各约同样操作量. 冻结 W 可省其参数梯度及状态, 却不能在前面有可训练层时省去 $\nabla_X L$.

27.4 逻辑长度总计 6000, 三条请求分别需要 4、8、12 页. 因此分页保留的位置数为

$$N_{\text{slots}} = 256(4 + 8 + 12) = 6144,$$

尾部浪费 $6144-6000=144$ 个槽位. 它只计独立尾块碎片，未计页表、预留块或前缀共享.

27.5 每参数平均 $\frac{4}{8}+\frac{4}{128}=0.53125$ 字节. 相对两字节权重为 $\frac{2}{0.53125}=\frac{64}{17}\approx 3.765$ 倍压缩. 逐行尾组、零点、对齐、未量化层和文件元数据都会增加实际尺寸；推理工作区另算.

27.6 参数权重 2 字节，梯度 4 字节，主权重和两个矩合计 12 字节，故 DDP 为 $18P$. 设均匀分片组 R，ZeRO-1 为 $(6+\frac{12}{R})P$，ZeRO-2 为 $(2+\frac{16}{R})P$，ZeRO-3 为 $\frac{18P}{R}$. 这些仅为持久状态；聚合和通信峰值未包含.

27.7 基线工作量 4.2×10^{21}，每卡有效速率 $2\times 10^{14}\times 0.3=6\times 10^{13}$. 期限 864000 秒，连续卡数 81.0185，向上取整 82；按 8 卡节点取整为 88. 需再次验证 88 卡拓扑能否保持所假定的 0.3 利用率.

27.8 按十进制单位，平均到达率 $\frac{400000}{1200}=333.33$ MB/s，大于 200 MB/s 服务率；每周期只能写 240 GB，积压增加 160 GB. 单检查点写完需 2000 秒，故长期不稳定；临时缓存只能延迟溢出，不能消除平均速率差.

27.9 可分别取上下文峰值翻倍、数据有效带宽减半、跨节点通信时间翻倍. 第一项增加 KV 并可能使注意力计算按长度平方增长，须重算工作量和峰值；第二项检查预取耗尽后的供给下界；第三项重算暴露通信时间. 若有效利用率来自包含相同数据等待和通信的端到端窗口，不再重复增加这两项；若仅测内核效率，则需单独建模. 报告每项口径和联合最坏情景，不能把三个独立最坏系数无条件相乘.

第二十八章　算力基础设施

28.1 寄存器允许 $\lfloor\frac{65536}{256\times 64}\rfloor=4$ 块，共享存储允许 $\lfloor\frac{64}{24}\rfloor=2$ 块，线程允许 8 块，块数限制 16，故最多 2 块. 共享存储改为 16 KiB 后允许 4 块，与寄存器共同成为主导，最多 4 块；实际分配粒度可能进一步降低驻留.

28.2 理想强度为 $I=\frac{B}{1+\frac{B}{2048}}$ FLOP/byte. 解 $I\geqslant 50$ 得 $B\geqslant\frac{102400}{1998}\approx 51.251$，最小整数 52. 此结论只排除理想带宽上界的阻挡；实际重复读取、矩阵单元效率、占用率和启动开销仍可能限制性能.

28.3 计算下界 0.02 秒，搬运下界 0.05 秒，可充分重叠时总下界为 0.05 秒. 只将操作量减半，计算项变 0.01 秒，总下界不变；只将流量减半，搬运项 0.025 秒，总下界降到 0.025 秒. 不能据此保证真实速度翻倍，因为原来是否重叠及其他开销尚未知.

28.4 可能路径为 NUMA1 内存，经跨插槽互连到 NUMA0 的 I/O，再传入 GPU；NUMA0 线程读取页面也需远端访问. 线程绑定不会自动搬迁既有页面. 可在 NUMA0

完成首次触页分配或迁移页，并绑定数据工作线程和合适的 I/O 路径；仍需测页分布、跨插槽流量与传输延迟，不可只看线程亲和性.

28.5 串行时间 $10+20+8+16=54$ 毫秒. 独立资源、足够缓冲下稳定间隔为 20 毫秒，首批仍约 54 毫秒. 读取和整理若争用同一不可重叠资源，应合成 30 毫秒阶段，间隔至少 30 毫秒；若只部分争用，应按共享资源占用率求界，而非继续取各阶段最大值.

28.6 $t=2(p-1)\alpha+\dfrac{2(p-1)m}{p\beta}$. 8 成员为 $14\alpha+\dfrac{1.75m}{\beta}$，64 成员为 $126\alpha+\dfrac{1.96875m}{\beta}$. 启动项增 9 倍，传输项也接近但不低于原值；此外拥塞和拓扑可能恶化，增加成员不是固定消息的免费加速.

28.7 取两个各有 32 台主机的机架，机架内端口均高速，但两架仅由一条慢链路连接. 端口带宽总和可以很高，跨机架任意流量仍受这一链路限制. 专家全互换可能令大量词元跨割传输，而局部相邻通信主要留在机架内. 可把高交互专家组放在同一高速域，并以跨割总字节率核算拥塞，不能靠端口总和判断.

28.8 上界为 $\min(\beta,\mathrm{IOPS}\times s)$. 4 KiB 时为 20000 KiB/s，即 19.53125 MiB/s；1 MiB 时请求能力允许 5000 MiB/s，但链路只有 1024 MiB/s，故上界 1 GiB/s. 并发可填满在途窗口，却不能突破请求处理器或传输链路上限.

28.9 全局 $\dfrac{2048}{90}\approx 22.756$ GiB/s，每节点 $\dfrac{256}{90}\approx 2.844$ GiB/s，另加校验、提交和训练流量. 分层暂存时，前台可在一致快照安全复制到暂存层后继续；远端持久化只有全部分片校验并原子发布清单后完成. 节点本地暂存丢失仍可能丢检查点，不能把前台恢复计算时刻记作远端完成.

28.10 清单包含运行/步骤编号、世界布局、64 个分片位置/字节数/摘要、模型优化器数据游标和随机状态版本. 分片先写临时命名空间，全部验证后发布唯一提交标志. 第 63 片后失效时该版本未提交，不得恢复；可续传或清理. 恢复发现任一摘要错误应拒绝整个版本，切换到前一已验证版本并报告数据丢失范围，不拼接不同步骤分片.

28.11 $w=\dfrac{C}{\tau}+\dfrac{\tau}{2M}+\dfrac{R_f}{M}$，导数 $-\dfrac{C}{\tau^2}+\dfrac{1}{2M}=0$，故 $\tau_*=\sqrt{2CM}$，二阶导正. M 缩至四分之一则间隔减半. 相关故障、非平稳失效率、异步保存竞争、长保存/恢复期间再次故障等会破坏近似，需改用实际失效与状态模型.

28.12 可能是激活重算增加无效算术、低效内核增加设备忙碌时间、通信轮询或重复失败重做、温度降频延长运行. 测实际能量积分与有效目标数、重算比例、内核时间、通信等待、频率/功率和恢复日志. 相同有效工作更久或功率更高均会升高单位能耗；忙碌率不能单独区分这些原因.

28.13 最少列设备/驱动/固件、框架与设备扩展、底层运行时、通信库、精度、模型修订、关键形状/算子和并行布局. 加载成功仅支持反序列化路径；算子正确需前反

向容差；通信可用需目标组、数据类型、次序和失效恢复；端到端达标还需质量、有效吞吐、峰值资源及恢复证据. 版本升级形成新矩阵并保留完整回退制品.

第二十九章　分布式训练

29.1　构造三个副本的微批计数为 $(1,3),(2,6),(4,4)$，对应每词元标量梯度为 $(0,2),(1,3),(4,0)$. 六个微批均值平均为 $\frac{10}{6}=\frac{5}{3}$；副本词元均值分别 1.5、2.5、2，副本均值平均为 2；全局目标和 $0+6+2+18+16+0=42$，目标数 20，正确值 2.1. 只有权重匹配各微批/副本有效计数，才能恢复全局词元目标.

29.2　令副本 r 目标梯度和为 S_r，全局有效目标数为 M. 平均归约产生 $R^{-1}\sum_r S_r$，再乘 $\frac{R}{M}$ 得正确梯度 $\sum_r \frac{S_r}{M}$. 也可每副本先乘 $\frac{R}{M}$ 再归约. 若框架额外除以累积步数 K，应补乘 K 或禁用该默认缩放；局部已除以各自计数时，必须先恢复目标和，不能只乘统一常数.

29.3　取两个等权副本标量梯度 10 和-1，阈值 1. 先裁剪后平均为 $\frac{1-1}{2}=0$；先平均为 4.5，再裁剪为 1. 非线性投影不能和求和交换. 应按声明的全局梯度范数完成归约及反缩放后裁剪.

29.4　近似存储 $A(\frac{L}{k}+k)$ 在 $k=\sqrt{L}=10$ 最小，结果 $20(10+10)=400$ MiB；相邻整数 9 或 11 都更大. 区间边界、残差分支、参数/梯度/优化器、通信桶、算子工作区和分配器余量均不在此简化量内，真实最优还受重算代价和不同层大小影响.

29.5　把 m 字节均分 R 块，每轮向下一成员发送一块并累加收到块，R-1 轮后每成员得到一块完整归约，发送量 $\frac{(R-1)m}{R}$. 全收集再经 R-1 轮传递归约块，合计启动 $2(R-1)\alpha$ 和传输 $\frac{2(R-1)m}{R\beta}$. 均匀链路和同步轮次是假设，不能代替拥塞测量.

29.6　$P_s=12, G_s=12, O_s=72$ GiB，R=4 时 DDP96、阶段 1 为 42、阶段 2 为 33、阶段 3 为 24 GiB. 另计 10 激活和 6 工作区，阶段 2 至少 49 GiB；阶段 3 若保留 4 GiB 聚合和 4 GiB 预取，则 44 GiB. 若设备可用 40 GiB，二者在此设定下均不够；需减少峰值、卸载或增加分片，不能只看持久状态 24 GiB.

29.7　沿用上游梯度 1，所有 ReLU 输入为正. $\nabla_{V_1}L=(1,2)^{\mathsf T}$，$\nabla_{V_2}L=(3,1)^{\mathsf T}$；$\nabla_{W_1}L=X^{\mathsf T}(1,2)=\begin{pmatrix}1&2\\2&4\end{pmatrix}$，$\nabla_{W_2}L=X^{\mathsf T}(-1,1)=\begin{pmatrix}-1&1\\-2&2\end{pmatrix}$. 按原列/行布局拼接等于完整矩阵微分，输入梯度为两成员贡献 $(1,2)+(-2,0)=(-1,2)$.

29.8　ReLU 反例：$\phi(1-2)=0$，而 $\phi(1)+\phi(-2)=1$. 列切分前馈层使每个隐藏坐标的完整预激活已在本地算出，因此可本地激活；随后行切分输出投影的部分和必须

归约. 允许的是不同隐藏坐标独立非线性，不是对同一坐标的未完成部分和分别非线性.

29.9 理想气泡比例 $\dfrac{p-1}{m+p-1}$，p=8 时要求 $\dfrac{7}{m+7}\leqslant 0.1$，得 $m\geqslant 63$. 增加微批会增加调度开销和在途激活，具体取决于 1F1B 等策略；若保持总批次不变，微批变小还可能降低矩阵效率，故 63 只是理想均衡模型下界.

29.10 对块 i 保存 (m_i,ℓ_i,u_i)，令 $m=\max_i m_i$，则 $\ell=\sum_{i=1}^{3}e^{m_i-m}\ell_i$、$u=\sum_{i=1}^{3}e^{m_i-m}u_i$，输出 $\dfrac{u}{\ell}$. 精确算术下均等于完整指数和，故可先合并任意两块. 全掩码块用空状态跳过；若三块全空则进入明确的全掩码行规则，不计算负无穷之差.

29.11 一个可行例为 DP=4、PP=4、TP=4，乘积 64. 固定 DP/PP 坐标的 4 成员做张量并行，固定 DP/TP 坐标的 4 成员沿层流水，固定 PP/TP 坐标的 4 副本做数据梯度归约. 全局有效词元只在四个 DP 副本及本次累积窗口上各计一次，不能把 TP 或 PP 重复表示乘进去. 物理上优先将高频 TP 组放入高速互连域，再验证流水和 DP 通信路径.

29.12 声明已把全部参数/优化器分片重排并验证摘要、消费位置和随机状态；固定全局有效目标归约及学习率时钟，调整微批/累积保持所需目标. 用同一小数据更新对比梯度与损失容差. 允许改变归约顺序时可声称语义一致和数值接近，不能声称逐位重放；若 DP 变化连带批量或样本顺序变化，应明确它是新训练日程.

第三十章　增量解码

30.1 约定生成 4 个词元后立即停止，不把最后输出再送入模型. 预填充处理提示位置 0 至 4，末端 logits 预测 y0，缓存 5；随后分别处理 y0 位置 5 预测 y1、y1 位置 6 预测 y2、y2 位置 7 预测 y3，缓存依次 6、7、8. 最终输出 4 个而缓存 8 位置；只有为继续生成再物化 y3 才到 9，输出长度不必等于已物化缓存长度.

30.2 嵌入与逻辑位置固定时历史层 0 状态相同. 若某层历史输入相同，其 QKV 相同；因果掩码使旧位置不读取新位置，注意力输出相同，逐位置归一化/前馈与残差也相同. 因此逐层历史 KV 不变. 双向注意力的旧行会读新增键，新分母和值改变，归纳在注意力这一步失败.

30.3 指数质量为 1、2、4，总量 7；输出 $(\frac{1}{7})(1,0)+(\frac{2}{7})(0,2)+(\frac{4}{7})(3,3)=(\frac{13}{7},\frac{16}{7})$. 漏自身时仍错误得到 $(\frac{1}{3},\frac{4}{3})$，相当于丢失 $\frac{4}{7}$ 的分布质量而重新归一，不是精度容差.

30.4 按最后输出不物化的约定，全量前向长度为 8、9、10、11，总线性位置 38；缓存为 8+1+1+1=11. 合法因果分数对：全量 36+45+55+66 = 202，缓存 36+9+10+11 = 66. 这是每层每头的逻辑配对，若内核计算被掩码的稠密元素，实际算术计数还会不同.

30.5 每词元字节为 $40 \times 16 \times (128+64) \times 2 = 245760$；乘 16384 得 4026531840 字节，即 3.75 GiB. 键值维不相等时不能用 $2d_k$ 替代二者和，分页元数据和预留另计.

30.6 固定层、长度、键值维及精度，缓存正比 KV 头数，4 头比 1 头为 4 倍. 32 个查询头仍各产生分数和输出，查询投影与前馈也未按 KV 头数缩减；广播和矩阵布局影响实际带宽，不能把缓存缩减 4 倍称为全模型计算缩减 4 倍.

30.7 若 $k_j = U_K c_j$，则 $q_t^{\mathsf{T}} k_j = q_t^{\mathsf{T}} U_K c_j = (U_K^{\mathsf{T}} q_t)^{\mathsf{T}} c_j$. $U_K \in \mathbb{R}^{d_k \times r_c}$，矩形矩阵不能与任意 $d_k \times d_k$ 旋转交换；例如 $R_j U_K$ 与 $U_K R_j$ 后者甚至维度不合法. 即使方阵也一般不交换，位置旋转须保持其真实分工.

30.8 逻辑位置 17 在第 $\lfloor \frac{17}{8} \rfloor = 2$ 个逻辑块，页表映射为物理块 7，偏移 1. 长度 19 共分配 24 槽位，尾部空 5 槽. 物理块编号 7 不意味着位置编码使用 7 或 57，应保留逻辑时间 17.

30.9 独立每请求长度 21，需 3 块，四条共 12 块；共享 16 前缀用 2 公共块，四条 5 词元尾部各 1 块，共 6 块. 独立分配 96 槽位、有效 84、尾碎片 12；共享分配 48 槽位、独特有效 $16 + 4 \times 5 = 36$，尾碎片仍 12. 节省来自重复前缀，不是所有尾部碎片被消除.

30.10 共享未满块若被一条分支直接写入会影响其他引用者，须分配私有块复制有效内容后追加. 引用归零只表示无活动所有者，设备内核可能尚未读完；需等完成事件或安全 epoch 后复用内存，否则产生异步读写竞态. 取消也必须按所有权只减一次.

30.11 全注意力删历史通常改变模型条件，是近似；原生滑窗删除未来不再可见的键值可保持其逐步计算语义；无引用前缀淘汰只降低未来命中率；活动抢占需换出完整状态或重算后恢复，影响成本和时延. 它们不能仅以一个 LRU 规则统一处理，须明确活动引用与可恢复边界.

30.12 键覆盖权重/结构摘要、适配器组合、精度布局、位置/掩码、完整词元前缀、模板处理器和图像特征依赖、共享命名空间. 仅温度作用于输出分布时已有 KV 可复用，但后续采样状态不同；适配器变化一般失效；模板若导致同一实际前缀且其他依赖相同可复用，否则失效；不同图像即便文本标记相同也不能复用其条件 KV.

第三十一章　批处理调度

31.1 TTFT 从到达到首词元，为 0.3 秒；首执行等待为 0.1 秒. 31 个输出间有 30 个间隔，$\mathrm{TPOT} = \frac{0.9 - 0.3}{30} = 0.02$ 秒. 总完成时间为 $0.3 + 30(0.02) = 0.9$ 秒. 等待已包含在 TTFT 内，不能再加一次；只有先扣除等待，才能把 TTFT 分成互不重叠阶段.

31.2　补充约定：已完成成员不继续参与计算，静态策略只禁止补入新成员．静态第一轮双序列耗 10 毫秒，B 完成；随后 A 独占三轮共 18 毫秒，A 在 28 毫秒完成；C 再耗 6 毫秒，在 34 毫秒完成，均值 $\dfrac{10+28+34}{3}=24$ 毫秒．连续策略第一轮 B 完成，第二轮 A 与 C 耗 10 毫秒，C 在 20 毫秒完成；A 再独占两轮，在 32 毫秒完成，均值 $\dfrac{62}{3}\approx 20.67$ 毫秒．连续策略改善均值但 A 稍晚；若静态内核仍计算填充槽，其轮时模型不同，须重新计数，不能沿用上述结果．

31.3　最大长度 256．线性利用率为 $\dfrac{64+64+256}{3\times 256}=\dfrac{1}{2}$；平方注意力利用率为 $\dfrac{64^2+64^2+256^2}{3\times 256^2}=\dfrac{3}{8}$．两式假定每条序列真实执行到同一填充长度．变长内核可跳过部分无效位置，且块对齐、掩码和工作区还有其他成本，因此这两个比例不是其实际加速比．

31.4　第 r 个新查询看见 $s+r$ 个键，故 $\displaystyle\sum_{r=1}^{q}(s+r)=qs+\frac{q(q+1)}{2}$．将长度 q 拆成 q_1,q_2 时，第二块历史变为 $s+q_1$，两块配对数之和仍等于原式．数学配对总数保持，而第二块需要重新读取历史键值，并多一次启动；配对守恒不能推出搬运或墙钟时间不变．

31.5　每页 16 位置，增长后长度为 259、257，页数为 $\lceil\frac{259}{16}\rceil+\lceil\frac{257}{16}\rceil=17+17=34$，较初始 33 页增加 1 页．第一条仍位于既有第 17 页，第二条跨过 256 边界．每词元 64 KiB 只描述有效数据增长，实际边际分配按整页发生，此处增加 1 MiB 而非按四词元只预留 256 KiB．

31.6　取两请求驻留区间 $[0,2)$ 和 $[1,3)$，观察窗 $[0,3]$．人数在三段依次为 1、2、1，面积为 $1+2+1=4$，也等于两请求各 2 的驻留时间之和．窗内平均人数 $\dfrac{4}{3}$、完成率 $\dfrac{2}{3}$、平均驻留 2，三者满足 Little 关系．存在窗外开始或结束请求时，有限窗恒等式使用区间交集长度；不能把完整驻留时间与截断人数面积混配．长期式需要边界残余除以窗长趋零．

31.7　记 $\rho=\dfrac{\lambda}{\mu}<1$．出生死亡平衡给 $\pi_{n+1}=\rho\pi_n$，归一化得 $\pi_0=1-\rho$、$\pi_n=(1-\rho)\rho^n$．由几何级数求导，$L=\dfrac{\rho}{1-\rho}$，服务中概率为 ρ，故 $L_q=\dfrac{\rho^2}{1-\rho}$．泊松到达看见稳态人数，条件于 n 个在系统者，总驻留为 $n+1$ 个参数 μ 指数变量之和；其拉普拉斯变换为 $\displaystyle\sum_n \pi_n[\frac{\mu}{\mu+s}]^{n+1}=\frac{\mu-\lambda}{\mu-\lambda+s}$，所以 $P(W>t)=e^{-(\mu-\lambda)t}$．连续批处理的多请求协同服务、依长度变化的轮时和调度优先级违反单服务台独立指数、FCFS 等前提．

31.8　不能．非负阶段 X,Y 满足 $\{X+Y>300\}\subseteq\{X>100\}\cup\{Y>200\}$，因而已知两个尾概率各不超过 0.01 时，$P(X+Y>300)\leqslant 0.02$；300 毫秒是一个 98% 覆盖

上界. 要保证 99%，可分别选择 p99.5 阈值再相加，或直接测量同请求的阶段和. 独立性也不能使两个分位点直接相加成为精确分位数.

31.9 设租户 A 每请求 10 词元，B 每请求 1000 词元，均持续积压. 每轮各完成一请求给予 B 约 100 倍输出份额；每轮各给 1000 词元则 A 完成 100 请求、B 完成 1 请求. 两者对应不同公平对象. 再令 B 具有更长历史，则同一输出词元还读取更多 KV，词元公平未必计算时间公平；可用经校准设备时间记账并配合最低服务份额，但成本估计误差必须监测.

31.10 净流入 $12-4=8$ KiB/s，填满时间 $\dfrac{64}{8}=8$ 秒. 上限到达后进入暂停生产或慢消费者状态，停止为其新增生成，继续排空并设置期限；若不能暂停或期限耗尽，则取消并释放其资源. 不得无界累积，也不得丢词元后仍声称完整流；终态与最后已确认事件号一并记录.

31.11 事件可为：请求 r 持有共享页 p 的一次引用；取消 CAS 将 r 置为终止；释放记录以 (r,p) 唯一键提交并将引用计数减一；迟到的抢占完成事件试图重新入队时因 r 已终止被拒绝；重复取消只读既有释放记录. p 只有全体引用归零且设备在途读写已通过完成事件或 fence 确认结束后回收，使用页代次防止旧事件释放复用后的新页. 终态单调、释放唯一键和页代次共同构成不变量，单靠队列中删除一次不够.

31.12 开放环按外部时钟发请求，能暴露积压增长；闭合环受前一响应限制，过载时到达率可能自动降低. 固定时长报告提交、接纳、成功、达标成功、拒绝、取消和未知数，分别以时间或对应请求集合为分母；记录全体已完成时延及未完成的删失/截止状态. 用新生成词元与总执行词元比识别重算，关联缓存换出事件；用输出缓冲占用、消费速率与慢客户端数识别背压. 不能只测成功请求平均速度而丢弃超时者.

31.13 在相同到达负载下关联队列年龄、预填充块时长、解码间隔及输出缓冲，使用开放环负载观察队列是否持续增长. 若排队增长且设备有效容量饱和，应限制准入并增加对应阶段容量；若长提示使解码间隔突增，应检查预填充分块和阶段预算，同时保留必要预填充份额；若生成完成但输出缓冲积压，应检查慢客户端、背压和取消传播. 显存余量只表示部分容量余量，不能证明计算或输出路径仍可满足期限. 扩容需要启动与预热时间，其间应按队列预算和期限执行排队或拒绝策略，并验证积压能够消退.

第三十二章　模型服务架构

32.1 权重相同只说明参数未改；模板改变特殊符号、角色边界和输入词元，从而改变 $p(y\mid x)$ 的实际条件. 因此它应形成新的服务制品修订，并冻结模板、分词器和生成配置. 可保留共同权重哈希用于去重，但接口兼容与质量回归需重新判断.

32.2 只允许非终态经 CAS 写入一个终态，取消和完成中先提交者获胜，另一方读回确定状态. 资源可能由异步清理者释放，故还须以请求/分配代次的唯一释放记录保护引用递减；终态 CAS 只约束状态字段，不能自动覆盖另外的内存操作. 迟到完成不得覆盖已取消状态，但远端业务结果仍需审计.

32.3 至少保留请求身份、制品与采样配置、已提交事件日志和单调序号、终态、恢复保留期限与身份授权. 若客户端已确认事件 10，重连从 11 继续；重复事件按序号去重. 事件重放读取已形成的结果，重新生成是新的随机计算，可能改变后文且重复成本，不能不加标记地冒充断点续传.

32.4 剩余预算 $2 - 1.2 - 0.3 = 0.5$ 秒. 再发起预计 1 秒的尝试不能满足原期限，应拒绝重试、缩小为可兑现的子操作，或返回期限不足. 所有下游必须接收原绝对截止时间或等价剩余预算；每重试重置 2 秒会破坏端到端约束.

32.5 启动信号允许加载和预热尚未完成；存活判断进程是否需要重建；就绪判断能否接收新请求. 过载通常应降低准入或就绪，不能把暂时排队当死锁反复重启. 重启会损失 KV 并引入冷启动，减少有效容量，形成更严重的过载反馈.

32.6 纯传输下界 $\dfrac{2}{40} = 0.05$ 秒，即 50 毫秒；30 请求/秒要求 60 GiB/s，超过单条 40 GiB/s 链路. 该下界不含握手、布局、排队与所有权提交. 需要增加有效带宽、减少迁移字节或限制接纳，不能因单请求 50 毫秒就认定系统可承载目标吞吐.

32.7 接收端先处于已准备但无输出权状态. 使用唯一请求代次和原子所有权记录确认提交结果；未知时查询权威记录并等待，不能双方凭超时自行继续. 只有持当前代次租约/隔离令牌的一端发布结果，旧拥有者被执行或输出网关拒绝. 数据复制成功不等于所有权迁移成功.

32.8 现有容量 1800、需求 2400 请求/分钟，恒定缺口 600/分钟，120 秒产生 1200 请求积压. 两种可行方法是预热容量池/预测提前扩容，以及受期限约束的准入和背压；也可降级到经验证的低成本路径. 若采用拒绝，积压公式应减去拒绝量，不能同时把拒绝请求算作仍排队.

32.9 影子复制请求而不影响正式响应，消耗额外资源，验证输出差异但不能完全模拟真实用户反馈；金丝雀让部分真实请求使用候选，需随机或明确定义分流及回滚；蓝绿保留完整两套环境，切换快速但容量高. 影子调用写工具会产生真实副作用，应只读、禁写或用明确隔离的业务验收环境，不能把“影子”当作副作用许可.

32.10 全局分布是按请求数加权的混合分布，不是各分位数的平均. 举例两个副本都固定分别耗 1 秒与 5 秒，只要 5 秒请求占比超过 1%，全局 p99 就是 5 秒；占比很小时可能为 1 秒. 应合并同口径直方图或逐例样本，保留计数权重和桶精度.

32.11 总尝试次数 N 满足 $E[N] = \displaystyle\sum_{j=1}^{4} P(N \geqslant j) = 1 + 0.2 + 0.2^2 + 0.2^3 = 1.248$，平

均重试数 0.248. 过载故障依赖共享资源，尝试相关且重试进一步增加负载；应采用总预算、退避抖动和熔断，而非据此认定只增加 24.8% 流量.

32.12 先记录故障前活动请求、页所有权与可服务副本；让指定工作进程失联，观察路由停止准入的时间、已输出前缀及唯一终态. 检查页释放或迁移代次、迟到事件、重复请求、未知结果及恢复后是否双重输出. 分别报告提交失败率、接纳后失败率、受影响请求恢复时长和剩余达标吞吐；不能把失联期间请求从统计分母删除.

第三十三章　推理计算优化

33.1 设 $m=\max(m_1,m_2,m_3)$，$\ell=\sum_i e^{m_i-m}\ell_i, u=\sum_i e^{m_i-m}u_i$. 代回各块定义后，分别等于全体有效位置的 $\sum e^{s_j-m}$ 与 $\sum e^{s_j-m}v_j$；这些有限和在精确算术中结合律成立，输出 u/ell 与合并次序无关. 浮点舍入仍可能造成微小次序差.

33.2 块 A 只有指数质量 1、值 0，块 B 只有指数质量 3、值 4. 正确输出 $\dfrac{1\times0+3\times4}{4}=3$；局部输出 0 和 4 等权平均为 2. 块权重必须正比全局指数质量，不能由块数或局部输出决定.

33.3 全掩码块仅表示本块对查询无贡献，可跳过；全掩码行表示所有块都无有效键，常规 Softmax 无定义. 用显式 empty 状态，首个非空块直接赋值，之后合并，避免 $-\infty-(-\infty)$. 全部空时按接口返回零/报错等明定规则，不能伪造概率归一化.

33.4 对一行 $p=\mathrm{softmax}(s)$，雅可比为 $\mathrm{diag}(p)-pp^{\mathsf T}$. 设概率上游 g_p，得 $g_s=p\odot(g_p-\langle p,g_p\rangle\mathbf{1})$. 注意力中 $g_{p,j}=g_o^{\mathsf T}v_j$，均值为 $g_o^{\mathsf T}o$；该减均值保证分数共同平移方向的梯度为零.

33.5 其余 80% 不变，总时间比例 $0.8+\dfrac{0.2}{10}=0.82$，加速 $\dfrac{1}{0.82}\approx1.2195$. 即使该算子无限快也不超过 1.25 倍；需确认优化没有改变其他时间项或引入额外开销.

33.6 补充构造数据：三个桶编译成本分别 2 秒、10 秒、1 秒，调用次数 1000、100、20，单次节省 3 毫秒、50 毫秒、10 毫秒. 总节省 3 秒、5 秒、0.2 秒，扣编译后净收益 1 秒、-5 秒、-0.8 秒，故仅第一个值得在这些调用量下编译. 缓存命中、内存和首请求期限也需纳入，调用量变化应重新求 $n\Delta t>C$.

33.7 p=(.1,.6,.3)，q=(.5,.2,.3)，条件接受率为 (.2,1,1)，接受质量 (.1,.2,.3)，总接受.6. 正残差 (0,.4,0)，拒绝后只出第二词元；加回拒绝质量后为 (.1,.6,.3)，恰为 p. 只有抽到 q 正支持的词元才计算 p/q.

33.8 $\sum_v\min(p_v,q_v)=\dfrac12\sum_v(p_v+q_v-|p_v-q_v|)=1-\dfrac12\|p-q\|_1$. q 为 0 而 p 正的词元不会从草稿抽中，但可经正残差产生；p 为 0 而 q 正的草稿必拒绝. 只要两者为同空间规范分布，精确残差法不要求相同支持.

33.9　同原始二元 logits 取 $(\log 4, 0)$，目标温度 1 给 p=(.8,.2)，草稿温度 2 给 q=$(\frac{2}{3},\frac{1}{3})$. 接受率为 (1,.6)，接受质量 $(\frac{2}{3},.2)$，拒绝质量 $\frac{2}{15}$ 补到第一词元，得到 p. 若误把草稿概率记成温度 1 分布，接受与残差就不再对应实际采样.

33.10　非终止且允许奖励词元时，$\mathbb{E}K = 1 + 0.9 + 0.9 \times 0.7 + 0.9 \times 0.7 \times 0.5 = 2.845$. 其中各接受率是给定前面已接受的条件概率；有 EOS 或剩余长度上限时截断相应项，不能仍用此无限续写表达.

33.11　第二候选拒绝意味着仅第一候选被确认，再从第二位置残差采样替代词元. 丢弃目标/草稿中第二候选及后续候选的投机状态，保留已确认前缀和第一候选. 替代词元先输出，但其 KV 若未物化需在下一轮先处理，或立即对齐两个模型缓存；逻辑输出长度与物化边界分别记录，不能保留被拒词元造成的隐藏后缀.

33.12　设被删质量 delta，保留归一均值 a、删除归一均值 b，则 $o = (1-\delta)a + \delta b$，截断输出 $o' = a$. 值范数 ⩽M 给 $\|o - o'\| = \delta\|b - a\| \leqslant 2\delta M$ （delta<1）. 多层传播、位置状态和采样可放大扰动，单行输出界不提供任务正确率保证.

33.13　以总时间除以总提交词元数衡量长期成本，方案 A 为 $\frac{12}{3} = 4$ 毫秒每词元，方案 B 为 $\frac{20}{4} = 5$ 毫秒每词元，故题设下应选 A. 接受率提高带来的提交量增长不足以抵消每轮成本增长. 统计时应以累计时间与累计提交数的比值估计，避免对不同提交数的单轮比值直接等权平均. 高并发下还需保持相同质量与采样协议，测量批量组成、长度分布、队列等待、设备利用率和尾延迟；草稿与验证争用资源后，单请求结论可能变化.

第三十四章　模型优化

34.1　三比特码 0 至 7，$s = \frac{1.8 + 0.7}{7} = \frac{5}{14}$，初始零点 $\frac{0.7}{s} = 1.96$，最近整数 z=2. 实际端点为 $-2s = -\frac{5}{7} \approx -.714286$ 与 $5s = \frac{25}{14} \approx 1.785714$. 整数零点量化使端点整体偏移；希望上界 1.8 必覆盖时需重新选择范围/尺度，不能称两端都精确命中.

34.2　若误差 e 在 $[-\frac{s}{2}, \frac{s}{2}]$ 均匀，$\mathbb{E}e^2 = s^{-1}\int_{-\frac{s}{2}}^{\frac{s}{2}} e^2\,de = \frac{s^2}{12}$. 若所有输入都在格点且无饱和，误差恒 0 而不是该值；该式是输入残差分布近似，不是量化器恒等式.

34.3　最近取偶数时 1.5→2、2.5→2、-1.5→-2、-2.5→-2；向零则 1.5→1、2.5→2、-1.5→-1、-2.5→-2. 向零有朝零的定向误差，最近舍入非平局时误差绝对值 ⩽s/2. 制品要固定舍入发生在加零点前还是后、平局规则、码域、饱和顺序和特殊值处理；浮点近半格点也需一致实现.

34.4 采用本章取偶数、码域 [-3,3]. 第二行恢复 $(\frac{40}{3}, 20, -\frac{40}{3})$，行 MSE 为 $\frac{200}{27}$；第一行逐张量为.02，逐行为 $\frac{1}{1350}$. 两行等元素数，整体逐张量为 $\frac{1}{2}(.02+\frac{200}{27}) = \frac{10027}{2700} \approx 3.713704$，逐行为 $\frac{1}{2}(\frac{1}{1350}+\frac{200}{27}) = \frac{10001}{2700} \approx 3.704074$. 改善来自小行，但大行误差主导；下游输入协方差和非线性另决定任务误差.

34.5 P=16777216, 四比特码 8388608 字节；每行 32 组，共 131072 组，元数据 524288 字节. 合计 8912896 字节 =8.5 MiB，十六位基线 32 MiB，压缩 $\frac{32}{8.5} = \frac{64}{17} \approx 3.7647$ 倍. 逐行已整除，本题忽略对齐及未量化参数.

34.6 最小化 $\frac{1}{2}\delta^{\mathsf{T}} H\delta$ 且 $e_i^{\mathsf{T}}\delta = d$. 拉格朗日式导数 $H\delta + \lambda e_i = 0$，得到 $\delta = \frac{dH^{-1}e_i}{(H^{-1})_{ii}}$，最小代价 $\frac{d^2}{[2(H^{-1})_{ii}]}$. 阻尼把 H 换 H+lambda I，改变目标及不可观测方向的代价；若 H 对角，各非 i 坐标补偿为 0. 实际 X 不相关要对应零交叉二阶矩而非口头 “不同特征”.

34.7 取 $H = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$、$w = (.6, .7)$，单位整数格点. 先量化第一坐标到 1，误差 +.4 令第二坐标减.2 至.5；按取偶数再到 0，最终 (1,0). 先量化第二到 1，误差 +.3 令第一减.15 至.45，再到 0，最终 (0,1). 固定坐标后不再变，故顺序可改变离散结果；本例已声明半格点规则.

34.8 对正对角 D，$WX = (WD)(D^{-1}X)$，保持实数函数. AWQ 主要通过激活重要性选择通道缩放来降低仅权重量化后的输出重建误差；SmoothQuant 平衡激活与权重动态范围，支持二者低精度. 恒等式相同不代表校准目标、搜索或执行格式相同，移动变换需保持全部共享分支一致.

34.9 $s_j = \frac{a_j^{\alpha}}{b_j^{1-\alpha}}$，alpha=.5 给 s=(8,1). 激活新范围 a/s=(8,4)，权重新范围 bs=(8,4). 两侧极值被平衡，乘积不变；这并未计算实际量化误差或证明该 alpha 最优，仍需代表性校准.

34.10 固定尺度和零点的阶梯函数在每个格区内导数 0，在跳点不可导，直接梯度几乎处处消失. STE 在反向指定近似导数如范围内 1、范围外 0，使优化能进行；它不是原离散函数的真实导数. 训练的舍入、饱和、粒度和部署整数路径应匹配，否则学到的误差补偿可能失效.

34.11 E4M3 为 4 指数 3 尾数，E5M2 为 5 指数 2 尾数（均另有符号位）；更多指数通常换更大动态范围，更多尾数通常换局部精度. 具体 NaN/无穷编码、指数偏置和次正规数决定最大值/下溢，乘法输入、累加和更新精度又是不同选择. 不能只看 FP8 名称推定范围、溢出行为或优化器状态.

34.12 $\sum_j s_x(q_{xj}-z_x)s_w(q_{wj}-z_w) = s_x s_w[\sum_j q_{xj}q_{wj} - z_w\sum_j q_{xj} - z_x\sum_j q_{wj} + nz_x z_w]$. 对称权重 $z_w=0$ 可去第二项和常数项，但激活非对称时仍保留 $-z_x\sum_j q_{wj}$. 分组 s_w 不同，要组内累加后各乘尺度再合并，不能把一个总尺度提出全部求和.

34.13 构造权重搬运仅占总时延 10%、其他计算/KV/排队占 90% 的长上下文负载. 即使前者理想加速 4 倍，总加速仅 $\frac{1}{.9+\frac{.1}{4}} \approx 1.081$；反量化开销可再抵消收益. 需分测权重和 KV 字节、算子与排队时间，不能由文件缩小四倍推导延迟四倍.

34.14 $o=\mathrm{softmax}(qK^{\mathsf{T}})V$，一阶有 $\delta o \approx J_{\mathrm{softmax}}(s)\,\delta s\,V + p\,\delta V$，其中 δs 含键误差与查询的内积. 只测值 MSE 覆盖第二项，漏掉键导致的注意力重分配及两误差相关性；不同查询/上下文还可放大相同逐元素误差.

34.15 保留集合 S 不改原数值时，平方误差为 $\sum_{i\notin S} w_i^2$，固定集合大小 k，交换论证 表明应保留绝对值最大的 k 项. 反例：w=(10,1)，输入总为 x=(0,1)，只保留一项时幅值法删第二项造成输出误差 1，而保留第二项输出误差 0. 权重范数目标不等于数据加权任务目标.

34.16 删除 w_q 产生 $\delta_q=-w_q$，Taylor 项为 $-g_q w_q + \frac{1}{2}H_{qq}w_q^2$（固定其他参数）. 一阶项可正可负，并可能压过二阶项. 训练接近结束不代表每坐标梯度为 0，随机批次、正则、早停及非平稳分布均可使其非零；应在明确校准目标下测量.

34.17 $H^{-1}=\frac{1}{5}\begin{pmatrix}2&-1\\-1&3\end{pmatrix}$. 删第一坐标 d=-1，补偿 $(-1,.5)$，新权重 (0,1.5)，代 价 $\frac{1}{[2(\frac{2}{5})]}=1.25$. 删第二坐标补偿 $(\frac{1}{3},-1)$，新权重 $(\frac{4}{3},0)$，代价 $\frac{1}{[2(\frac{3}{5})]}=\frac{5}{6}$. 该驻点二次目标下删第二项更便宜；非零一阶梯度需重新解.

34.18 列向量约定下 $U,G\in\mathbb{R}^{d_f\times d}$、$D\in\mathbb{R}^{d\times d_f}$. 保留中间集合 S 时取 $U[S,:], G[S,:], D[:,S]$；上投影/门控偏置取 S，输出偏置维度不变. 梯度、Adam 矩、主权重和分片索引同步裁剪，残差输出 d 不改；若使用行向量实现应转置对应索引.

34.19 稠密 2P 字节，非零比例 rho 的数值 + 索引为 4rho P，节省要求 rho$<\frac{1}{2}$. 行指针、对齐和稀疏元数据会收紧阈值；不规则访问、缺少匹配内核和低矩阵单元利用率仍可使稀疏计算更慢.

34.20 总平方范数 126. 秩 1 误差 $\sqrt{26}$、相对 $\sqrt{\frac{26}{126}}\approx .4543$；秩 2 为 $\sqrt{10}$、相对约.2817；秩 3 为 1、相对 $\frac{1}{\sqrt{126}}\approx .08909$. 若矩阵 m×n，因子预算需 $r(m+n)$，可先找达到误差阈值的最小 r，再检查预算；未给 m/n 和预算时无法断言哪一秩可行.

34.21 任意秩 $\leqslant$r 近似 A 的列空间投影 P 满足 $\|W-A\|_F^2 \geqslant \|(I-P)W\|_F^2$. 设 WW^{T} 特征值为 σ_i^2，秩 r 正交投影最多捕获前 r 个特征值之和，故误差至少谱尾和；取前 r

奇异向量即达界. 若截断处 $\sigma_r = \sigma_{r+1}$，等特征子空间中可选择不同基/子空间而同样最优，矩阵近似可能不唯一.

34.22 原参数 $4096 \times 11008 = 45088768$，和为 15104. 需 $r < \dfrac{45088768}{15104} \approx 2985.2203$，整数 r⩽2985. r=512 比例 $\dfrac{7733248}{45088768} = \dfrac{59}{344} \approx 17.1512\%$；r=2048 比例 $\dfrac{30932992}{45088768} = \dfrac{59}{86} \approx 68.6047\%$. 这是参数节省条件，不是两次小 GEMM 一定更快.

34.23 普通秩 1 保留 4、删除 1，输出误差期望为 $1^2 \times 25 = 25$. 加权问题看 $WC_x^{\frac{1}{2}} = \mathrm{diag}(4,5)$，最优保留第二方向，映回近似为 diag(0,1)，误差为 $4^2 \times 1 = 16$. 输入高方差方向可以推翻仅按权重奇异值的选择.

34.24 $q = \mathrm{softmax}(\dfrac{z_S}{\tau})$，$p$ 为冻结教师，$\dfrac{\partial[\tau^2 \mathrm{KL}(p\|q)]}{\partial z_S} = \tau(q-p)$. 高温下 $q_i \approx \dfrac{1}{K} + \dfrac{z_i - \bar{z}}{K\tau}$，故梯度近似师生中心化 logit 差/K. 共同平移 logits 不改变概率，未经中心化的欧氏 logit 差会惩罚一个不可辨识方向.

34.25 本章定义 lambda 乘软项，故 $\lambda = .8$ 时损失 $.2\log 2 + .8(4D) \approx .755412$，梯度 $.2(-.5,.5) + .8(-.6,.6) = (-.58,.58)$. 硬标签改第二类时硬梯度 (.5,-.5)，软梯度不变，总为 (-.38,.38)：两项竞争而此设定软项占优；硬损失仍 log2，因为学生均匀.

34.26 第一，词表事件不同，不能直接把同索引概率当同一类别；第二，文本被分成不同位置数，条件前缀也不同. 需要显式词表/跨度映射或改用序列级教师文本. 序列蒸馏保留采样出的可读答案与任务结构，但丢失未采样候选的完整概率和不确定性，受教师采样偏差影响.

34.27 实数乘积 $(10A)(\dfrac{B}{10}) = AB$. 独立量化后含 $10AE_{B'} + \dfrac{E_{A'}B}{10} + E_{A'}E_{B'}$，格点、动态范围和尺度元数据未必按反向比例等变，误差可能改变. 若理想量化器和尺度恰能完全同比缩放，也可能保持同误差；题目强调可能性而非必然.

34.28 固定数据/版本、质量门槛、硬件、长度/并发与采样；核对模型文件、权重/KV/工作区峰值和实际内核路径；预热同步后测首词元、逐词元、吞吐及尾延迟，并与未压缩基线比较. 恢复重载和任务/安全回归也须通过. 两个局部数字不能证明时延、总显存或质量达标，需报告误差范围和不支持时的回退路径.

34.29 单项验收不能推出组合达标. 各步骤会改变后续步骤的输入分布和权重结构，误差还会跨层传播，因此不能将独立测得的退化简单相加. 固定原始模型为共同参考，记录每一步制品，按实际处理顺序重新校准并评估最终组合；中间制品间的比较用于定位新增误差，同时保留相对原模型的累计质量差异. 对有依据的候选顺序进行受控比较，使用独立任务与失败切片评估，并核对目标设备的算子支持、峰值内存和端到端时延. 只有最终制品在既定质量与资源约束下通过验证，才支持部署.

第三十五章　双向语言建模

35.1　任选句对 $(\mathrm{CLS}, a, \mathrm{SEP}, b, \mathrm{SEP})$，词元编号为各词表 ID，位置编号为 $(0,1,2,3,4)$，常见句段编号为 $(0,0,0,1,1)$．句段编号只选择一条可学习向量加入表示，未产生负无穷屏蔽；只要注意力 mask 允许，两个句段仍可互读．实际特殊词元 ID 以固定制品为准．

35.2　第 1 键无效后，双向剩余分数指数为 2、3，权重 $(0, \frac{2}{5}, \frac{3}{5})$，输出 $(\frac{2}{5})4+(\frac{3}{5})10 = 7.6$．因果还要屏蔽第 3 键，仅第 2 键可见，权重 $(0,1,0)$，输出 4．必须在剩余允许集合重新归一化，不能仅将原权重置零．

35.3　独立选择时空集合概率为 $(1-.15)^{20} = .85^{20} \approx .038760$．固定选 3 个则空集合概率为 0，每个位置仍有边缘概率 $\frac{3}{20}$，但不同位置的联合选择概率为 $\frac{3\times 2}{20\times 19}$，不等于 $(\frac{3}{20})^2$；指示量负相关．两种策略具有不同批量监督量分布．

35.4　条件于位置已被选中，显式保留分支概率为 0.1，随机替换分支恰好抽回原词元概率为 $\frac{0.1}{V}$，总和 $0.1+\frac{0.1}{V}$．此处假设 MASK 符号不等于原词元且随机词表含原值．若统计所有候选位置还需加上未被选中的概率，不能把两个条件空间混用．

35.5　逐序列平均为 $\frac{.2+.8}{2} = .5$；逐监督词元平均为 $\frac{1\times .2+3\times .8}{4} = .65$．前者给短序列每个目标更大权重，后者给每个监督词元相同权重．应在训练目标和报告指标中固定同一分母，并排除空监督序列的除零情形．

35.6　把监督位置收集为 $G \in \mathbb{R}^{N_M\times d}$，目标独热矩阵 Y，平均损失的分数梯度为 $D = \frac{Q-Y}{N_M}$．由 $U = GE^{\mathsf{T}} + b_V$ 得 $\overline{E}_{\text{out}} = D^{\mathsf{T}}G$．损坏输入也通过查表使用 E，因此总梯度还加按实际损坏后编号散射求和的输入梯度；输出目标是原词元，两条索引路径不同．

35.7　例如起点分数 $(0,5,1)$，终点分数 $(5,0,1)$，独立 argmax 选起点 2、终点 1 而非法．按 $i \leqslant j$、$j-i+1 \leqslant 2$ 枚举合法片段，分数依次 $(1,1):5, (1,2):0, (2,2):5, (2,3):6, (3,3):2$，最优为 $(2,3)$．采用起终分数之和并固定并列规则；若模型定义另有长度惩罚，目标需同步改变．

35.8　同值两序列各为 $.9^2 = .81$，异值各为 $.1^2 = .01$，总质量 $2(.81+.01) = 1.64$．归一化后，同值条件概率变成 $\frac{.81}{.81+.01} = \frac{81}{82} \approx .987805$，并非原来的.9．归一化能构造一个联合分布，但不保证保留原先条件预测．

35.9　先固定原文、候选位置、数据隔离、模型初始状态和训练预算；分别记录掩码率、随机替换规则和有效目标数．比较参数质量时，在共同固定的验证损坏集合及相同下游任务上评价；比较任务难度时，固定模型交叉评估两种损坏策略．若每次验证随机重新抽位置，需多次重复并报告波动，不能把抽样难度变化误当改进．

第三十六章　混合专家模型

36.1　选中专家 1、2. 选中内权重为 $(\frac{4}{7},\frac{3}{7})$，输出 $(\frac{4}{7},\frac{6}{7})$；保留全专家概率时为 (.4,.3)，输出 (.4,.6)．后者权重总和.7，缺少的.3 没有被重分配，所以两种输出不同；选择集合相同不代表函数相同.

36.2　固定集合、选中归一化时 $y=\sum_i w_i F_i$，对选中 logit 的梯度为 $w_j g^{\mathsf{T}}(F_j-y)$. $k=1$ 时 $w_j=1, y=F_j$，因此为零．若保留全专家 Softmax 概率，$y=p_jF_j$，p_j 仍受所有 logit 影响，故通常有非零梯度．选择边界本身不可微，局部零梯度不表示改变排名永远不改变输出.

36.3　两个选中专家均输出 v 时，归一化混合始终为 v，$F_j-y=0$，路由任务梯度为零．但专家输出的上游梯度分别是 w_jg，专家参数通过其雅可比仍可得到非零梯度. 路由选择无差别不等于专家停止学习；辅助均衡也可能继续更新路由.

36.4　取 $N=2$、$k=1$、$E=2$、每专家容量 $C=1$，则 $EC=Nk=2$．若两个词元都选择专家 1，按词元编号接收一个、拒绝另一个．路径拒绝率 $\frac{1}{2}$，受影响词元率 $\frac{1}{2}$，完全失去专家处理率 $\frac{1}{2}$．专家 2 的空闲槽位不会自动把已选路径变为合法替代；重新路由是另一策略.

36.5　专家 2 的四个权重为 $\frac{2}{9},\frac{1}{3},\frac{7}{18},\frac{2}{9}$．按降序为词元 3、2，然后词元 1 与 4 并列；题设补充并列按词元编号升序，故仍拒绝词元 4，输出与正文的同序接收例相同．若并列优先较大编号，则拒绝词元 1：不重归一化时 $y_1=(\frac{7}{9})x_1$，其余恢复原无拒绝输出．该例中的是否改变依赖明确的并列规则；不能凭方法名称声称一定改善.

36.6　设 $L=E\sum_i f_i\bar{p}_i$，$\bar{p}_i=N^{-1}\sum_t p_{ti}$，并对离散统计 f 停止梯度．由 Softmax 导数得 $\frac{\partial L}{\partial z_{tj}}=(\frac{E}{N})p_{tj}(f_j-\sum_i f_i p_{ti})$．两个批次大小 N_1,N_2 时，正确全局概率为 $\frac{N_1\bar{p}^{(1)}+N_2\bar{p}^{(2)}}{N_1+N_2}$；直接各占一半将小批次词元过度加权，路由计数也须在同一统计范围聚合.

36.7　一个无偏置门控专家含三矩阵，共 $3dd_f=3145728$ 参数．16 个路由专家总计 50331648，每词元选 2 个执行 6291456．新增 2 个同尺寸共享专家后，总量 56623104，激活量 12582912．这里只数专家矩阵；若计入路由器还需加 $dE=8192$，注意力、归一化及输出头另计.

36.8　d_f 减半、E 和 k 加倍保持 $3kdd_f$ 及专家总矩阵参数近似不变，但路由 logit 数量翻倍，选择/分发元数据和路径数增多．每专家矩阵更窄、每路径批次更小，内核利用率可能下降；通信消息数量、负载尾部与合并开销也变化．需同硬件同 token 预算测量，不能由乘加次数守恒推出时延守恒.

36.9 若分发权重 $D_{ts}=\dfrac{\exp z_{ts}}{\sum_u \exp z_{us}}$，则每个槽位列和为 1；合并权重 $C_{ts}=\dfrac{\exp z_{ts}}{\sum_r \exp z_{tr}}$，则每个词元行和为 1. 两词元一个槽位、权重各 $\frac{1}{2}$ 时，槽输入 $\dfrac{x_1+x_2}{2}$，若专家为恒等，则第一个输出直接含 x2. 仅在最终合并处加因果 mask 不能消除已经混入的未来信息.

36.10 低并发使每专家输入行数很少，矩阵乘法退化为低利用率的小矩阵；大量专家权重可能仍需从显存读取，计算量低但带宽占主导. 跨设备每轮小消息暴露固定通信延迟，路由偏斜又使少数专家成为串行尾部. 应记录逐专家批量、读取字节、通信耗时和请求尾延迟，再决定合批、放置或卸载策略.

36.11 需要逐层专家数/容量/共享布局、总驻留权重精度、路由和 KV 状态、峰值工作区、设备分片及卸载方案；还需请求长度并发、批次分布、溢出与重试、内核和通信数据. 以实际候选模型在目标硬件测吞吐、首词延迟和尾延迟，并验证任务质量. 平均负载会掩盖热点，激活参数也不等于显存.

第三十七章　长上下文建模

37.1 长度比 $s_L=\dfrac{32768}{4096}=8$，末位置 32767 映为 4095.875；端点比 $s_E=\dfrac{32767}{4095}$，映射 $f(i)=\dfrac{4095i}{32767}$，末位置恰为 4095. 两者差 0.875 个参考位置，均可作为明确约定，但不能在缓存、新查询或不同层间混用，否则相对相位不一致.

37.2 原相位 $\Delta\omega=\dfrac{2048}{256}=8$，因子四插值后为 2. 固定 q、k 时精确旋转范数界为 $2\|q\|\|k\|\left|\sin(\dfrac{2-8}{2})\right|=2|\sin 3|\|q\|\|k\|$，还可用较松的 $6\|q\|\|k\|$. 该界仅约束当前二维点积分量；不能直接约束深层网络，因为 q、k 也可能随扩展改变.

37.3 若 $b'=b\,s^{\frac{d_r}{d_r-2}}$、$\omega_k=b^{-\frac{2k}{d_r}}$，则 $\dfrac{\omega'_k}{\omega_k}=s^{-\frac{2k}{d_r-2}}$. 最高频 $k=0$ 比例 1，最低频 $k=\dfrac{d_r}{2}-1$ 比例 $\dfrac{1}{s}$. $d_r=2$ 时分母为零，且仅一个频带，无法同时满足“最高频不变”和“最低频缩为 1/s”的非平凡要求.

37.4 设目标分数 a、其余分数 c，权重 $\dfrac{1}{[1+(T-1)e^{c-a}]}=.9$，解得 $a-c=\log[9(T-1)]$. T 由 101 至 901 需要增加 $\log(\dfrac{900}{100})=\log 9\approx 2.197225$. 真实模型的干扰分数不相等，多头值内容、层间表示和位置编码也变化；本构造揭示分母稀释，不是通用长度性能公式.

37.5 范围为 $1+6(8-1)=43$ 个连续位置，远离序列起点时最早为 $i-42$. 一条信息路径依次为输入 $i-42$、第 1 层 $i-35$、第 2 层 $i-28$、第 3 层 $i-21$、第 4 层

$i-14$、第 5 层 $i-7$、第 6 层 i. 每步最多向后传七格. 存在路径不保证信息在有限维状态中不被干扰或压缩损失.

37.6 不能. 因果图的边从不晚于查询的位置指向该查询，位置 0 只能读取位置 0；由层数归纳，任一层的位置 0 都不含未来输入. 让后续位置读取该全局键只传播初始摘要，不会使它反向汇总全文. 可将摘要位置放在片段末端或显式维护外部记忆，但需重新定义协议.

37.7 令短序列概率 p，要求 $2048p = 16384(1-p)$，得 $p = \frac{8}{9}$，长序列概率 $\frac{1}{9}$. 词元份额相等不代表独立样本量或重复率相等；若某桶可用原始序列很少，有放回抽样会重复使用同一内容，必须记录唯一来源与曝光次数.

37.8 本章每序列每词元裸 KV 为 128KiB. 保留上下文与最大生成总长 $65536 + 8192 = 73728$，三条并发需要 $73728 \times 3 \times 128\,\mathrm{KiB} = 27\,\mathrm{GiB}$. 未含模型权重、临时激活和工作区、分页碎片、元数据、调度预留，以及可能的额外推测路径. 还要核对模型允许的总长度，不只看输入上限.

37.9 可构造两条证据：“项目甲批准预算 70 万元”“补充批复增加 12 万元”，问题为最终批准额，答案 82 万元；移除任一必须证据形成不可确定变体. 独立操纵总长度、两证据相对位置和同格式干扰文件数量，其余文本分布固定. 评分同时检查数值、单位、两条来源及无答案拒答. 只测最长输入开头会漏掉中部遗漏、跨远距离合并和干扰鲁棒性.

37.10 摘要便于总览，但可能删除金额、例外和单位；应保留逐条来源及摘要到原文映射. 检索节约上下文，却可能漏掉跨条款定义或补充协议，需版本一致的检索与关联展开. 直接长窗口保留更多原文，但仍有位置遗漏和预算限制. 可用检索定位加原文精确抽取、单位校验与交叉核对，保留原始条款、查询、候选、计算中间值和不确定项；不可仅依赖自由摘要结算.

第三十八章　开源模型体系

38.1 把列向量记号转置，$c = hD^{\mathsf{T}} \in \mathbb{R}^{1\times r}$、$k_i = cU_i^{\mathsf{T}} \in \mathbb{R}^{1\times d_k}$、$v_i = cV_i^{\mathsf{T}} \in \mathbb{R}^{1\times d_v}$. 分数 $q_i k_i^{\mathsf{T}} = (q_i U_i)c^{\mathsf{T}}$，吸收查询 $q_i U_i$ 为 $1\times r$. 输出 $\sum_j a_j c_j V_i^{\mathsf{T}} = (\sum_j a_j c_j) V_i^{\mathsf{T}}$，先汇总 r 维潜变量再投影到值宽度.

38.2 $U^{\mathsf{T}} R_j^{\mathsf{T}} R_t q$ 随 j 变化，一般需逐历史位置计算，不能当作同一个当前查询. 若 U 为同一二维子空间的标量单位矩阵，则与旋转可交换，但旋转角度仍依赖 j；“可交换”本身不等于“可消除位置”. 不使用旋转、所有相对旋转恰为同一矩阵，或相关向量位于旋转恒等子空间时，才可在该限定范围复用.

38.3 MQA 每层每词元存 $2 \times 1 \times 64 = 128$ 个元素，潜在形式仍为 $128 + 32 = 160$ 个元素. 因此该构造下潜在形式为 MQA 的 1.25 倍，反而更大. 它仍可能带来其他结构质量或算子取舍，但不能由方法名称保证缓存占优；比较需固定键值宽度、潜在维数和位置分支.

38.4 驻留权重为 $40 \times 10^9 \times 2 = 80 \times 10^9$ 字节. 4×10^9 只描述某词元选择的计算参数，不表示其他专家已经从显存消失. 8GB 只能近似所选参数的两字节体积；若需要降低驻留量，必须另有分片、卸载或量化策略，并计入相应通信和质量成本.

38.5 冻结权重、输入事实、任务和解码预算，逐层比较模板序列化后的角色、特殊词元、编号及停止协议，再比较首步 logits 和完整输出. 以官方适配模板作为一个有来源的对照，而非假定其对任何任务最优. 若两模板表达不同条件，质量差异首先属于协议实验；只有控制该变量后才能归因模型本体. 工具解析与媒体占位也需同步检查.

38.6 $L_a = 0$ 时 $M = b_c L_s d_k d_v$，在给定有限状态模型中与 T 无关；$L_s = 0$ 时 $M = 2 b_c L_a T h_{kv} d_h$，退化为线性增长的 KV. 固定状态把大量历史压缩到有限维度，可能发生信息干扰与不可区分历史，不能保证任意长依赖精确恢复. 还须考虑数值精度和状态更新误差.

38.7 检查分词器有效 ID 最大值、added tokens、模板控制符号、是否为硬件对齐预留行、输入输出权重是否共享及头部形状. 核对权重和配置提交、适配器目标版本及转换历史. 若最大有效 ID 小于行数且多余行有明确策略，可为合法填充；若新 ID 未配训练参数或配置错配，则需修复. 仅凭 V 与 V+16 不作唯一结论.

38.8 每条任务同时记录模式、正确/有效完成、内部与可见总生成词元、工具调用轮数、端到端时延、重试次数及停止原因. 固定任务集、模型提交、模板、工具环境，并预登记两模式相同总预算或各自预算下的比较问题. 分别报告质量与成本分布及超时失败，不能只在成功回答中计算平均时延.

38.9 开放题没有唯一模型选择. 可选择一个具有官方仓库的开放权重检查点，例如 Falcon 或 Pythia 系列中的具体版本，清单包含：精确 repo 与 commit；逐层结构和权重头；Base/Instruct 及训练目标的报告依据；tokenizer、模板、EOS 和工具协议；模态及 processor；许可原文版本；分片摘要、转换和适配关系. 把“模型卡声称支持某长度”标为声明，把实际配置字段标为静态观察，把任务正确率/后端吞吐标为待测. 未实际读取时不能填造参数与许可证；完整性标准是六维均有来源或明确未知.

38.10 MLA 改变注意力历史状态的表示及相关投影路径，应按实际实现计算潜在缓存、附加位置状态、投影计算及缓存读取成本. MoE 改变前馈层的专家选择与执行，应分别统计全部专家权重、每词元激活专家、路由缓冲及跨设备派发和归并通信. 长上下文更突出缓存随长度增长的成本；专家并行还受到路由负载不均衡和网络开销

制约. 每词元激活参数量描述单次执行涉及的参数，全部专家仍需按放置或卸载方案存储，运行时还需要缓存与工作空间，因此不能仅以激活参数估算部署显存.

第三十九章　视觉 Transformer

39.1　网格 $G_h=\dfrac{96}{8}=12, G_w=\dfrac{128}{16}=8$，补丁数 $N=96$，块内元素 $F=3\times8\times16=384$. 零起始最后补丁 $n=95$ 对应网格 $(11,7)$；按块内通道优先、再行列的约定，最后元素 $f=383$ 对应 $(c,u,v)=(2,7,15)$，映回图像 $(c,h,w)=(2,95,127)$，批索引可为 0 或 1. 若通道后置，需同时改索引公式与权重排列.

39.2　无重叠整除时，每像素 h,w 唯一分解为 $h=g_hP_h+u, w=g_wP_w+v$，其中商余数唯一；再将 (g_h,g_w) 和 (c,u,v) 编码为 n、f，构成双射，逆映射按商余数还原. 重叠会复制像素，逆需检查一致并累加/归一；忽略边界丢信息所以不可逆；填充需保留原尺寸并裁掉新增位置，才能恢复原图.

39.3　令块内置换矩阵为 P，行向量展平改成 $X'=XP$，取 $W'_E=P^TW_E$，则 $X'W'_E=XPP^TW_E=XW_E$. 交换两个像素等于交换 W 对应两行. 只改 X 不改 W 会让不同像素被原位置权重相乘，例如左右差核可能算成上下或混合差，形状虽不变而输出语义改变.

39.4　设抽取算子 $X=UI$，投影 $Z=XW$，则 $D_I=U^T(D_ZW^T)$. 展开为某像素的梯度等于所有包含该像素的窗口、对应块内位置以及输出通道贡献之和. 重叠时同像素出现多次，单纯逆重排只保留一次会漏梯度；若为重建图像求平均，要明确这是另一归一化算子，而反向应为前向抽取算子的转置.

39.5　线性映射由 12×5 矩阵定义，其秩至多 5. 秩零化度定理给核空间维数 $12-\operatorname{rank}W\geqslant7$，所以存在至少七维非零扰动 δx 满足 $\delta xW=0$. 这说明线性编码不能保留全部输入差异，却不能单独判断分类性能，因为被丢弃方向可能与标签无关；任务充分性需要数据与分类目标证据.

39.6　令置换 $P=\operatorname{diag}(1,P_v)$ 固定 CLS，对无位置的自注意力有 $Q'=PQ, K'=PK, V'=PV$，行 Softmax 满足 $\operatorname{softmax}(PSP^T)=P\operatorname{softmax}(S)P^T$，输出为 PO；逐位置前馈、归一化和残差也等变. 由层归纳 CLS 行不变，分类不变. 加入位置后同时置换内容和位置仍只是整行置换；只置换内容则改变内容与空间的配对，一般改变输出. 推理关闭随机丢弃或同步其置换是该证明的条件.

39.7　沿用源网格 $\left[\begin{smallmatrix}0&2\\4&6\end{smallmatrix}\right]$，取角点对齐，连续场为 $f(y,x)=4y+2x$. 目标 y 取 $0,\dfrac{1}{3},\dfrac{2}{3},1$，x 取 $0,\dfrac{1}{2},1$，结果为 $\left[\begin{smallmatrix}0&1&2\\\frac{4}{3}&\frac{7}{3}&\frac{10}{3}\\\frac{8}{3}&\frac{11}{3}&\frac{14}{3}\\4&5&6\end{smallmatrix}\right]$. 中心对齐会使用另一坐标和边界处理，结果不同. CLS 表示汇聚角色，没有二维像素位置，单独保留其参数，不能插入网格混合.

39.8 主要线性项 $(4+2\rho)Ld^2$ 与注意力项 $2L^2d$ 相等，正 L 下约去 $2Ld$ 得 $L=(2+\rho)d=6\times512=3072$. 这是乘加量交点；内存搬运、融合、并行度、矩阵利用率和工作区不同，真实时间交点必须测量，不能据此直接决定内核选择.

39.9 固定补丁尺寸提高分辨率，块内维数 F 与投影 W 不变，位置数及注意力激活增大，学习位置表需重采样. 固定分辨率减小补丁，F 和 W 形状改变、位置数增大，因此不能只插值位置表就复用全部权重. 共享编码层的 d 不变时其参数可保留；两种变化都改变输入尺度分布，质量需验证.

39.10 有效掩码 $m_i\in\{0,1\}$，$n_v=\sum_i m_i>0$，读出 $g=\sum_i \frac{m_i h_i}{n_v}$. 给定上游 D_g，$D_{h_i}=\frac{m_i D_g}{n_v}$；填充位置为零. 若除以含填充的 N，则 g 和梯度被缩为 $\frac{n_v}{N}$，让表示幅度与填充比例相关. 全空样本必须拒绝或单独定义，不能除零.

39.11 小字符裁剪可能删掉决定标签的字，镜像可能改变文字含义，Mixup 软标签隐含组合目标而非真实混图类别概率. 可固定按原始图/文档分组的数据划分，在相同模型、训练步数和预算下分别比较无增强、语义保持裁剪、翻转和 Mixup. 测试使用固定处理器，并按字符大小与方向切片；同一原图变体不得跨训练测试泄漏.

39.12 先用原始样本与处理器日志核对数值范围和归一化，再核对 RGB/BGR 与通道轴；用可手算网格验证 patchify 及位置索引；用保存中间激活核对 CLS/GAP 和掩码；最后核对分类头列与标签 ID 映射. 每步以固定输入对照参考输出，定位第一个偏离处. 形状检查只能排除维度错误，不能排除数值尺度、位置或标签语义错误.

第四十章　多模态建模

40.1 对行式投影 $Z_{n,j}=\sum_{c,u,v} I_{c,g_hP_h+u,g_wP_w+v}W_{(c,u,v),j}+b_j$，令卷积核 $K_{j,c,u,v}=W_{(c,u,v),j}$、步幅等于块大小、无填充即可逐项相等. 增加填充改变边界像素与输出网格，重叠改变同像素被使用次数，步幅不等于块大小时不再是简单无重叠分割；仍可按相同窗口定义建立抽取与卷积的关系.

40.2 两边各增 1.5 倍，面积和视觉词元数增 $1.5^2=2.25$ 倍；若只计视觉位置且无额外特殊词元，全自注意力位置对增 $2.25^2=5.0625$ 倍. 加入 CLS 或固定文本前缀后应计算 $\frac{(2.25N+c)^2}{(N+c)^2}$，不能仍称精确 5.0625 倍. 编码器重采样等额外开销另计.

40.3 行损失 $\mathcal{L}_r=B^{-1}\sum_i[-S_{ii}+\log\sum_j e^{S_{ij}}]$ 导数为 $\frac{R_{ij}-\delta_{ij}}{B}$. 列损失同理为 $\frac{C_{ij}-\delta_{ij}}{B}$；两者平均得 $G_{ij}=\frac{R_{ij}+C_{ij}-2\delta_{ij}}{2B}$. 行、列分母分别归一同图不同文与同文不同图，除特殊对称数据外不相同.

40.4 $\tau = 0.5$ 时 S 对角为 2，$p = \dfrac{e^2}{1+e^2} \approx 0.880797$. 平均损失 $\log(1+e^{-2}) \approx 0.126928$，逻辑值梯度对角 $-\dfrac{1-p}{2} \approx -0.0596015$、非对角为正 0.0596015. 若对数缩放 $a = \log(\dfrac{1}{\tau})$，$\dfrac{\partial \mathcal{L}}{\partial a} = \sum G_{ij}S_{ij} = -0.238406$. 表示梯度还乘 $\dfrac{1}{\tau} = 2$；降低温度同时改变概率与链式缩放，不是把原梯度简单翻倍.

40.5 非零向量 a，$u = \dfrac{a}{r}, r = \|a\|$，微分 $du = \dfrac{da}{r} - \dfrac{a(a^T da)}{r^3}$，故 $J = \dfrac{I - uu^T}{r}$. $u^T J = 0$，输出梯度在球面切空间，径向变化被去除. 零向量处方向未定义，需明确带 epsilon 的归一化实现及其真实导数；不能继续使用 $r = 0$ 的该式.

40.6 取一图的两个正文本 1、2 及一个负文本 3，预测 $(0.8, 0.1, 0.1)$. 均匀正对交叉熵目标 $(0.5, 0.5, 0)$，梯度为 $(0.3, -0.4, 0.1)$，提升较弱正对. 正对总概率损失 $-\log(p_1+p_2)$ 的梯度为 $p_j - \dfrac{\mathbf{1}_{j\in P} p_j}{0.9}$，即约 $(-0.0889, -0.0111, 0.1)$，提升两正对的总质量但不要求均衡. 二者对应不同学习目标.

40.7 Q 为 0 时两个逻辑值均 0，A 为 $(\dfrac{1}{2}, \dfrac{1}{2})$，输出 $Z = (1, 2)$. 一般值梯度 $D_V = A^T D_Z$，即每行 $\dfrac{D_Z}{2}$；本题沿用目标 $Z^* = (1, 2)$，因此 $D_Z = 0$，值及其余梯度全零. 平均来自此次相同逻辑值，不是任意查询与键下重采样的固定行为.

40.8 取固定 $A = (\dfrac{1}{2}, \dfrac{1}{2})$，$M = 1 < N_v = 2$，任意非零向量 v 使扰动 $\Delta V = (v, -v)^T$ 满足 $A\Delta V = 0$. 一般 A 的秩至多 M，故每值通道有至少 $N_v - M$ 维核. 真实注意力 A 可能随输入变化，该固定核扰动不一定保持 A；结论只证明局部固定线性值映射有信息压缩，不证明整个非线性网络对所有此类扰动均不敏感.

40.9 正确配对为位置 4 输出预测 A_1，位置 5 输出预测 A_2，位置 6 输出预测 EOS. 标签数组按目标位置存储后由因果损失移位，不能再重复移位. 提示位置本身不作为目标，却产生第一个答案的预测；删除所有提示位置输出会丢掉 A_1 监督. 视觉位置仍应被注意力读取，即使其直接目标被忽略.

40.10 路径为损失到答案逻辑值、穿过冻结语言主干的输入雅可比、到连接器输出及连接器参数. 冻结只令语言参数不更新，不能把主干输出从计算图 detach，否则连接器得不到损失梯度. 若视觉编码器也冻结，可把其输出视为常量，但连接器及下游输入梯度仍需记录或重算；低内存方案不能删掉必要链式导数.

40.11 样本数 $3 \times 16000 = 48000$；不补齐窗口数 $1 + \lfloor \dfrac{48000 - 400}{160} \rfloor = 298$. 若两端各补 200 样本且沿相同步长，数量变为 $1 + \lfloor \dfrac{48000}{160} \rfloor = 301$，但实际工具可能另有边界约定，应核对. 时间中心和有效观测长度也随补齐变化，不只是张量多三行.

40.12 随机相位下，长度 d 不超过采样间隔 Δ，命中相位区间长度 d，所以 $P(\text{命中}) = \dfrac{d}{\Delta} = \dfrac{0.2}{0.8} = 0.25$. 它是帧包含事件的概率，识别正确还须乘相应条件概率；漏采事件

不能由提升识别器直接恢复. 相位非均匀或采样自适应时需重新建模.

40.13 以参考时间 t 表示设备时间 $t' = at + b$，用至少两个分离的同步锚估计 a、b，多锚可拟合并检查残差；校正 $t = \dfrac{t' - b}{a}$. 窗口 $[t_0, t_1]$ 中心可代表事件位置，但编码器若读取到 t_1，最早可用时刻不能早于 t_1 加处理延迟. 时钟突跳或非线性漂移时分段拟合，不能一条直线覆盖全部会话.

40.14 计数任务可保持背景、对象外观不变，仅改变对象个数，并标注正确整数；声画任务可保持语音内容但改变时间偏移. 加入缺图、静音或错误模态对照，指标用计数准确率、绝对误差、匹配正确及同步误差，并以原场景分组. 模型输出改变只证明敏感，可能对水印或噪声敏感；需变化方向与正确答案一致才支持利用模态信息.

第四十一章　扩散模型的概率原理

41.1 递推条件方差 $v_t = \alpha_t v_{t-1} + \beta_t, v_0 = 0$.若 $v_{t-1} = 1 - \bar{\alpha}_{t-1}$，则 $v_t = \alpha_t - \bar{\alpha}_t + 1 - \alpha_t = 1 - \bar{\alpha}_t$. 展开独立噪声之和另得 $\sum_{s=1}^{t} \beta_s \prod_{j=s+1}^{t} \alpha_j$；每项等于 $\prod_{j=s+1}^{t} \alpha_j - \prod_{j=s}^{t} \alpha_j$，望远镜相消为 $1 - \bar{\alpha}_t$. 独立与单位协方差是方差直接相加的前提.

41.2 设 $E[x_0] = m, \mathrm{Cov}(x_0) = \Sigma$ 且噪声独立，$x_t = a_t x_0 + b_t \epsilon$，其中 $a_t = \sqrt{\bar{\alpha}_t}, b_t = \sqrt{1 - \bar{\alpha}_t}$. 于是均值 $a_t m$、协方差 $a_t^2 \Sigma + b_t^2 I$. 只有 m 为 0 且 $\Sigma = I$ 时二阶矩保持标准形式；即便满足二阶矩，原分布非高斯时也不能据此断言任意 t 都是标准高斯.

41.3 共享总噪声构造 $x_t = a_t x_0 + b_t \epsilon$ 给条件跨时协方差 $b_s b_t I$. 马尔可夫链在 $s < t$ 时条件协方差为 $\sqrt{\dfrac{\bar{\alpha}_t}{\bar{\alpha}_s}}(1 - \bar{\alpha}_s) I$，通常不同. 每个时刻二者仍都是均值 $a_t x_0$、方差 $b_t^2 I$ 的高斯；同边缘不确定联合分布. 画噪声趋势可用共享耦合，但不能把它当作独立单步噪声轨迹.

41.4 设 $u = x_{t-1}$. 负对数中的二次项为 $\dfrac{\|x_t - \sqrt{\alpha_t} u\|^2}{2\beta_t} + \dfrac{\|u - \sqrt{\bar{\alpha}_{t-1}} x_0\|^2}{[2(1 - \bar{\alpha}_{t-1})]}$. 精度 $A = \dfrac{\alpha_t}{\beta_t} + \dfrac{1}{1 - \bar{\alpha}_{t-1}} = \dfrac{1 - \bar{\alpha}_t}{[\beta_t (1 - \bar{\alpha}_{t-1})]}$，线性项 $h = \dfrac{\sqrt{\alpha_t} x_t}{\beta_t} + \dfrac{\sqrt{\bar{\alpha}_{t-1}} x_0}{1 - \bar{\alpha}_{t-1}}$. 配方为 $\dfrac{A \|u - \frac{h}{A}\|^2}{2}$ 加常数，故方差 A^{-1}、均值 $\dfrac{h}{A}$. t 为 1 时前一状态就是给定 x_0，为退化点质量；此处应使用重建项而非正方差高斯 KL 公式.

41.5 对 d 维 $q(x_T \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_T} x_0, (1 - \bar{\alpha}_T) I)$ 与标准高斯，KL 为 $\frac{1}{2}[\bar{\alpha}_T \|x_0\|^2 - d\bar{\alpha}_T - d \log(1 - \bar{\alpha}_T)]$. 由一般高斯 KL 的均值平方、迹和对数行列式项代入即可. 再对数据平均时 $\|x_0\|^2$ 换成其期望. 高维或大尺度会放大残余均值项，日程末端 $\bar{\alpha}_T$ 不够小时不能默认终端与标准先验已接近.

41.6 写 $p_\theta(x_0) = E_{q(x_{1:T}|x_0)}[\frac{p_\theta(x_{0:T})}{q(x_{1:T} \mid x_0)}]$，由负对数凸性得负对数似然不超过该比值负对数的期望. 把 q 反向分解为 $q(x_T \mid x_0)\prod_{t=2}^{T} q(x_{t-1} \mid x_t, x_0)$，p 分解为 $p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$. 逐项合并得到终端 KL、$t = 2 \dots T$ 后验与反向核 KL 的期望，以及 $-E \log p_\theta(x_0 \mid x_1)$ 重建项. 重要性比需有支持包含条件.

41.7 固定反向方差时均值误差项为 $\frac{\|\mu_\theta - \tilde{\mu}_t\|^2}{2\sigma_t^2}$，噪声参数化给误差系数

$$\frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}.$$

取 $\sigma_t^2 = \beta_t$ 后权重为 $\frac{\beta_t}{[2\alpha_t(1-\bar{\alpha}_t)]}$. 若预测方差，KL 中的 $\mathrm{tr}(\Sigma_\theta^{-1}\tilde{\Sigma})$、$\log\det\Sigma_\theta$ 及均值的逆方差权重均依赖参数，不得丢弃而仍声称等价似然训练.

41.8 要估计内部项总和 $\sum_{t=2}^{T} E[\ell_t]$，任选正概率 π_t 且总和 1，采 t 后用 $\frac{\ell_t}{\pi_t}$，期望即原和. 例如 $\pi_t \propto w_t$ 可能减少权重差异，但最优方差还依赖损失二阶矩. 若目标为时间平均，另除 $T-1$；每元素归一则另除 d. 两种归一改变常数尺度或跨形状权重，不能漏掉后仍称同一目标.

41.9 对固定 x_t, t 令 $m = E[\epsilon \mid x_t, t]$，展开 $E[\|\epsilon - f\|^2 \mid x_t, t] = E[\|\epsilon - m\|^2 \mid x_t, t] + \|f - m\|^2$，交叉项因条件均值为零消失，故最优 f 为 m. 多个 x_0, ϵ 可能产生同一 x_t，条件方差不可消除；预测期望不等于识别本次真实噪声.

41.10 $a^2 + b^2 = 1$，$x_t = ax_0 + b\epsilon, v = a\epsilon - bx_0$ 构成正交矩阵 $\left[\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right]$，逆为 $x_0 = ax_t - bv, \epsilon = bx_t + av$. 固定 x_t 时，$\delta x_0 = -(\frac{b}{a})\delta\epsilon$（a 非零），$\delta\epsilon = a\delta v$、$\delta x_0 = -b\delta v$. 所以等权噪声损失等于 SNR 乘 x_0 误差平方，也等于 a^2 乘 v 误差平方；等权 v 损失等于噪声误差平方/a^2. 端点零系数处必须用不奇异参数化，而非直接相除.

41.11 跨步 $t \to s < t$ 的噪声方差为 $\sigma_{t\to s}^2 = \frac{\eta^2(1-\bar{\alpha}_s)(1-\frac{\bar{\alpha}_t}{\bar{\alpha}_s})}{1-\bar{\alpha}_t}$. 故 $\eta = 0$ 为零，$\eta = 1$ 为右侧去掉 η^2 的值；相邻步退化为后验方差. 确定性更新仍以随机 x_T 为输入，其推前分布可以非退化，不能因每一步无新噪声就说只生成一个固定样本.

41.12 原噪声误差 0.1 时 KL 为 $1.25(0.1)^2 = 0.0125$. 误差幅度 0.2 时均值误差幅度 $\frac{0.2\times 0.2}{\sqrt{0.8}\sqrt{0.28}} \approx 0.08452$，KL 为 $1.25(0.2)^2 = 0.05$，变四倍；符号改变不影响此同方差 KL. 两步日程末端 $\bar{\alpha}_2 = 0.72$ 仍保留大量信号，计算只验证给定后验误差，不证明能由标准高斯生成真实数据.

第四十二章　条件视觉生成

42.1　单头视觉 $H_v \in \mathbb{R}^{N_v \times d}$，文本 $H_t \in \mathbb{R}^{N_t \times d_t}$，投影得 $Q \in \mathbb{R}^{N_v \times d_k}, K \in \mathbb{R}^{N_t \times d_k}, V \in \mathbb{R}^{N_t \times d_v}$．分数和权重为 $N_v \times N_t$，输出 $N_v \times d_v$，多头再拼接投影．交叉配对数 $N_v N_t$；联合注意力以 $N_v + N_t$ 位置构成 $(N_v + N_t)^2$ 对，包含双方内部及反向交互，结构不同不能只按成本等价替换．

42.2　固定潜图面积、无补齐且不计特殊位置，块边从 4 到 2 使每块面积降为四分之一，位置数增 4 倍，完整分数矩阵增 16 倍．逐位置线性层通常随位置数增 4 倍，补丁输入投影维数也改变；VAE 编码、解码、文本编码等不按 16 倍变化．实际速度还取决于分块内核和带宽．

42.3　扰动 $x_t = a_t x_0 + b_t \epsilon, b_t > 0$ 时条件核得分为 $\nabla_{x_t} \log q(x_t \mid x_0, c) = -\dfrac{x_t - a_t x_0}{b_t^2} = -\dfrac{\epsilon}{b_t}$．在可交换微分积分且密度正的条件下，对后验 $x_0 \mid x_t, c$ 求期望得 $\nabla \log p_t(x_t \mid c) = -\dfrac{E[\epsilon \mid x_t, c]}{b_t}$．b 为 0 时核退化、该除法不成立，必须通过端点极限或单独边界处理．

42.4　本章约定 $\epsilon_g = \epsilon_u + w(\epsilon_c - \epsilon_u) = w\epsilon_c + (1-w)\epsilon_u$．比较系数可得 $w = 1 + s$，s 为 3 时 w 为 4．不同库把 guidance 称为 s 或 w，参数同名不能直接照搬；条件缺失训练、数值精度与双路推理协议也需一致．

42.5　在固定 t 上，$s_u + w(s_c - s_u)$ 可写成 $\nabla \log[p_t(x \mid c)^w p_t(x)^{1-w}]$，但这只是该时间局部得分形式．各 t 的幂次密度未必满足采用的前向/反向动态所需的一致边缘方程，且归一化与模型误差另有条件．因此不能据此断言采样终点恰为对应幂次数据密度，更不能保证 w 越大越真实．

42.6　$q(z \mid x) = \mathcal{N}(\mu, \operatorname{diag}\sigma^2)$ 到 $\mathcal{N}(0, I)$ 的 KL 为 $\frac{1}{2}\sum_j (\mu_j^2 + \sigma_j^2 - 1 - \log \sigma_j^2)$，来自均值项、协方差迹、维数项与行列式比．写 $z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I)$ 后，随机源独立参数，可通过 μ, σ 对重建损失反传．重参数化不消除采样方差，也不保证潜空间无信息损失．

42.7　原元素数 $1024^2 \times 3 = 3145728$，潜元素 $128^2 \times 16 = 262144$，比值 12．空间面积比 64，但通道从 3 到 16，所以元素比为 $\dfrac{64 \times 3}{16} = 12$．元素个数还没有乘精度字节数，不能直接等同文件压缩率或端到端加速．

42.8　令 U 为训练条件路径速度，固定 (z, t, c) 时平方回归同样分解为不可约条件方差加 $\|u - E[U \mid z, t, c]\|^2$，最优场为条件期望．对光滑紧支撑 φ，若微分可交换，$\dfrac{dE[\varphi(Z_t)]}{dt} = E[\nabla\varphi(Z_t)^T u(Z_t, t, c)]$，等价弱式 $\partial_t p + \nabla \cdot (pu) = 0$．从弱式到唯一 ODE 运输还需速度正则性、可积性等条件，不能只凭训练损失低得出精确生成．

42.9 补足初值和区间为 $z(0)=1, t\in[0,1]$，步长 $h=\dfrac{1}{4}$. 递推 $z_{k+1}=1.25z_k$，四步为 $1.25^4=2.44140625$；精确 $e\approx 2.71828183$，低估约 0.27687558. 局部截断 $O(h^2)$、固定区间全局 $O(h)$ 需光滑与稳定条件；网络近似误差不包含在该纯求解器例子中.

42.10 F 帧每帧 S 位置，完整配对 F^2S^2；空间后时间分解约 FS^2+SF^2. 后者先在同帧或同位置混合，跨帧跨位置关系通过多层/组合间接传播，通常不等于一次完整注意力. 还须比较布局变换、同步、局部性、表达能力及质量，FLOPs 相同或较少不能证明性能与能力优势.

42.11 例如“把杯子改为红色，背景不动”，分别度量目标掩码内颜色/语义符合率和掩码外像素误差、结构差异. 使整图偏红可提高粗图文相似度却损害背景保持；强制每个外部像素不变又可能留下杯边不自然. 应声明边界羽化范围和允许光照变化，并用人工或任务标注检验，不能只用一项总相似度.

42.12 清单含主干、VAE、文本编码器/分词器、潜通道/尺度、时间映射、预测对象、采样网格、CFG 约定、控制模块和精度. 换 VAE 须核对潜维数、缩放、分布与解码；换文本编码器须核对维数、模板及训练语义空间；换求解器须核对时间方向、预测类型和端点. 组件可加载只证明接口部分兼容，仍需固定输入的数值及生成质量验证.

42.13 仅改变步数会改变数值离散误差；训练插值路径为直线，也不能推出模型边缘速度沿实际轨迹恒定或单步 Euler 精确. 采用一致性训练或蒸馏时，应核对模型预测对象、噪声或时间端点、条件输入与对应采样协议，确保单步执行符合训练目标. 使用固定测试条件和覆盖多个随机种子的样本集合，比较质量分布、多样性、条件遵循及失败切片，并计入编码、解码和其他流水线阶段测量实际延迟及资源占用. 记录模型制品和求解配置；单张成功样本或 FLOPs 下降不足以支持整体部署结论.

第四部分

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of ACL*. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162/.

[2] Ashish Vaswani et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems* 30 (2017). arXiv: 1706.03762. URL: https://arxiv.org/abs/1706.03762.

[3] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. May 29, 2023. arXiv: 2305.18290. URL: https://arxiv.org/abs/2305.18290.

[4] Zhihong Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. Feb. 5, 2024. arXiv: 2402.03300. URL: https://arxiv.org/abs/2402.03300.

[5] Edward J. Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[6] Hugging Face. *Transformers*. Hugging Face. URL: https://github.com/huggingface/transformers (visited on 09/08/2026).

[7] Hugging Face. *PEFT*. Version 0.20.0. URL: https://github.com/huggingface/peft/tree/v0.20.0 (visited on 09/11/2026).

[8] Hugging Face. *TRL: SFT Trainer*. Version 1.13.0. URL: https://github.com/huggingface/trl/blob/v1.13.0/docs/source/sft_trainer.md (visited on 09/11/2026).

[9] LLaMA-Factory contributors. *LLaMA-Factory*. URL: https://github.com/hiyouga/LlamaFactory (visited on 09/11/2026).

[10] Axolotl maintainers and contributors. *Axolotl: Open Source LLM Post-Training*. URL: https://github.com/axolotl-ai-cloud/axolotl (visited on 09/11/2026).

[11] Samyam Rajbhandari et al. *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*. Oct. 4, 2019. arXiv: 1910.02054. URL: https://arxiv.org/abs/1910.02054.

[12] DeepSpeed contributors. *DeepSpeed*. URL: https://github.com/deepspeedai/DeepSpeed (visited on 09/11/2026).

[13] PyTorch contributors. *PyTorch: FSDP2 fully_shard Implementation*. Version 2.11.0. URL: https://github.com/pytorch/pytorch/blob/v2.11.0/torch/distributed/fsdp/_fully_shard/_fully_shard.py (visited on 09/11/2026).

[14] NVIDIA. *Megatron-LM*. URL: https://github.com/NVIDIA/Megatron-LM (visited on 09/11/2026).

[15] Hugging Face. *Accelerate: DeepSpeed Integration*. Version 1.15.0. URL: https://github.com/huggingface/accelerate/blob/v1.15.0/docs/source/usage_guides/deepspeed.md (visited on 09/11/2026).

[16] vLLM contributors. *vLLM: Quickstart*. Version 0.29.0. URL: https://docs.vllm.ai/en/v0.29.0/getting_started/quickstart/ (visited on 09/11/2026).

[17] SGLang contributors. *SGLang: Inference Framework*. Version 0.5.19. URL: https://github.com/sgl-project/sglang/tree/v0.5.19 (visited on 09/11/2026).

[18] NVIDIA. *TensorRT-LLM*. URL: https://github.com/NVIDIA/TensorRT-LLM (visited on 09/11/2026).

[19] SGLang contributors. *SGLang: Server Arguments*. Version 0.5.19. URL: https://github.com/sgl-project/sglang/blob/v0.5.19/python/sglang/srt/server_args.py (visited on 09/11/2026).

[20] SGLang contributors. *SGLang: Sending a Request*. URL: https://docs.sglang.io/docs/basic_usage/send_request (visited on 09/11/2026).

[21] Tri Dao et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. May 27, 2022. arXiv: 2205.14135. URL: https://arxiv.org/abs/2205.14135.

[22] Yaniv Leviathan, Matan Kalman, and Yossi Matias. *Fast Inference from Transformers via Speculative Decoding*. Nov. 30, 2022. arXiv: 2211.17192. URL: https://arxiv.org/abs/2211.17192.

[23] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019. URL: https://aclanthology.org/N19-1423.

[24] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. June 19, 2020. arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239.

习题解析与实验指导