序号
1
2
3
4
5
6
7
8
9
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#
#

409	数据质量评估常用的自动指标包括（ ）。
410	Instruction Template 在微调预处理阶段常做的操作包括（ ）。
411	样本标注一致性评估方法包括（ ）。
412	数据探索阶段常用的可视化手段包括（ ）。
413	数据集融合时常见的字段映射方式包括（ ）。
414	训练前对长文本预处理时常采用的策略包括（ ）。
415	微调样本中常见的不良样本包括（ ）。
416	推理参数 temperature 的典型作用是（ ）。
417	top_p（核采样）的典型作用是（ ）。
418	模型推理效果对比实验中常记录的指标包括（ ）。
419	提升推理效果的常见做法包括（ ）。
420	A/B 测试在模型推理效果对比中常关注（ ）。
421	推理优化常用方法包括（ ）。
422	对比实验中保证结论可信的常见做法包括（ ）。
423	推理服务常见的瓶颈包括（ ）。
424	推理对比时控制随机性的做法包括( )。
425	评估生成质量常见的人工维度包括（ ）。
426	常见的模型量化格式包括（ ）。
427	4-bit 量化常见的精度损失来源包括（ ）。
428	部署格式选型时常关注的因素包括（ ）。
429	GGUF 格式的典型特点是（ ）。
430	AWQ 量化的核心思路包括（ ）。
431	ONNX 部署的常见优势包括（ ）。
432	TensorRT 部署的典型优化包括（ ）。
433	模型量化与部署取舍时常见的权衡维度包括（ ）。
434	在受限显存上部署 LLM 的常见手段包括( )。
435	部署后的推理引擎选型中常见考虑包括（ ）。
436	GPU 显存管理的常见做法包括（ ）。
437	CUDA 环境检查的常用命令包括（ ）。
438	Python 依赖管理的常见做法包括（ ）。
439	降低显存占用的常见技巧包括（ ）。
440	模型加载时显存优化的常用手段包括（ ）。
441	环境依赖检查时常看的版本字段包括（ ）。
442	多卡训练时的显存均衡做法包括（ ）。
443	包管理工具选型时常见考虑包括（ ）。
444	在 Linux 服务器上准备 GPU 推理环境的常见步骤包括（ ）。
445	显存 OOM 时的常见排查思路包括（ ）。
446	综合评估报告中常包含的内容包括（ ）。
447	确保实验可复现的常见做法包括（ ）。
448	常见的自动评估指标包括（ ）。
449	提交结果时通常需要附上的内容包括（ ）。
450	评估指标设计时常考虑的维度包括（ ）。

451	实验设计中常见的对照设置包括（ ）。
452	人工评估常见的做法包括（ ）。
453	写评估报告时常见的数据可视化方式包括（ ）。
454	提交最终结果前常见的自检步骤包括（ ）。
455	评估多个模型时常用的统计检验方法包括（ ）。
456	LoRA 训练中合理选择 target_modules 的常见位置包括（ ）。
457	QLoRA 训练中的关键组件包括（ ）。
458	LoRA 中影响最终表达能力的常见超参数包括（ ）。
459	LoRA 训练前的常规准备包括（ ）。
460	对 LoRA adapter 进行保存与加载的常用做法是（ ）。
461	LoRA 中 lora_alpha 与 lora_rank 的关系影响（ ）。
462	QLoRA 相比 LoRA 在资源受限场景下的优势包括（ ）。
463	下列哪些是 PEFT 库支持的常见微调方法？（ ）
464	LoRA 训练过程中常见的回调配置包括（ ）。
465	LoRA 中可以并行训练多个 adapter 实现多任务学习的常见做法包括（ ）。
466	RAG 中常见的长文档切分方法包括（ ）。
467	向量检索中常见的相似度度量包括（ ）。
468	在 RAG 中使用 Rerank 模型的好处包括（ ）。
469	选择 RAG 的 Embedding 模型时常关注的指标包括（ ）。
470	RAG 系统常见的人工评估维度包括（ ）。
471	典型 RAG 系统的链路构成包括（ ）。
472	在 RAG 系统中拼接 Prompt 时常见的注意事项包括（ ）。
473	向量检索常见的索引结构包括（ ）。
474	RAG 文档切片时为减少语义断裂常采用的策略包括（ ）。
475	RAG 中常用的检索评估指标包括（ ）。
476	KV cache 的主要作用是（ ）。
477	在并发推理服务中常见的高吞吐方案包括（ ）。
478	Flash Attention 相比普通 Attention 的优势包括（ ）。
479	动态批处理的常见实现要点包括（ ）。
480	主流推理引擎 vLLM 的核心特性包括（ ）。
481	常见的推理参数及其作用包括（ ）。
482	Self-Attention 优化的常见方向包括（ ）。
483	高并发推理服务的常见瓶颈包括（ ）。
484	KV cache 的常见变体包括（ ）。
485	采样参数中影响多样性的包括（ ）。
486	客服数据清洗的常规步骤包括（ ）。
487	客服数据隐私脱敏的常见处理对象包括（ ）。
488	客服对话去重时需要考虑的维度包括（ ）。
489	客服数据增强常用的方法包括（ ）。
490	客服样本类别不均衡处理方法包括（ ）。
491	客服多轮对话数据常见的格式标准包括（ ）。
492	客服数据中常见需要过滤的异常情况包括（ ）。

493	数据集融合时常见的处理环节包括（ ）。
494	客服对话样本质量的人工抽检指标包括（ ）。
495	对敏感字段做脱敏时常采用的策略包括（ ）。
496	Few-shot Prompt 设计的常见要素包括（ ）。
497	Chain-of-Thought（CoT）的典型形态包括（ ）。
498	ReAct 提示词的关键字段包括（ ）。
499	System Prompt 中通常包含的设定包括（ ）。
500	Instruction Template 设计时常包含的核心字段包括（ ）。
501	Prompt 调优过程中常用的方法包括（ ）。
502	结构化 Prompt 设计中常见要素包括（ ）。
503	Few-shot 示例选择的原则包括（ ）。
504	CoT 在推理任务中常通过以下方式激发（ ）。
505	为防止系统提示被忽略可采用的做法包括（ ）。
506	监督微调数据常见的数据格式标准包括（ ）。
507	padding 与 truncation 在微调数据预处理中的作用包括（ ）。
508	数据切分（train/val/test）时需要遵循的原则包括（ ）。
509	数据质量评估常用维度包括（ ）。
510	微调数据预处理阶段 Instruction Template 拼接的常见做法包括（ ）。
511	样本标注过程中常用的质检手段包括（ ）。
512	数据探索阶段常用的描述统计包括（ ）。
513	数据集融合时常见的处理步骤包括（ ）。
514	微调数据预处理的常规流程包括（ ）。
515	训练前的数据过滤常见规则包括（ ）。
516	推理参数 temperature 的常见取值与效果对应包括（ ）。
517	top_p 和 top_k 采样的常见特点包括（ ）。
518	评估生成模型效果时常记录的指标包括（ ）。
519	模型推理效果调优的常见方向包括（ ）。
520	A/B 测试在模型上线流程中常见的做法包括（ ）。
521	推理服务侧的常见优化手段包括（ ）。
522	对比实验中为保证公平常见的做法包括（ ）。
523	推理服务常见的瓶颈包括（ ）。
524	评估 LLM 生成效果时常见的人工维度包括（ ）。
525	针对幻觉问题常见的调优手段包括（ ）。
526	4-bit 量化常见的精度损失来源包括（ ）。
527	GGUF 格式典型支持的特性包括（ ）。
528	GPTQ 与 AWQ 的常见区别包括（ ）。
529	选择部署格式时常见的考虑因素包括（ ）。
530	ONNX 部署的常见优势包括（ ）。
531	TensorRT 部署中常见的优化技术包括（ ）。
532	模型部署中常见的格式包括（ ）。
533	推理引擎选型时常见的对比维度包括（ ）。
534	评估量化模型效果时常用的手段包括（ ）。

535	GPTQ 量化过程的关键步骤包括（ ）。
536	GPU 显存管理的常见做法包括（ ）。
537	CUDA 环境常见检查项包括（ ）。
538	Python 依赖管理的常见工具包括（ ）。
539	降低显存占用的常见技巧包括（ ）。
540	模型加载时显存优化的常用手段包括（ ）。
541	环境依赖检查时常看的版本字段包括（ ）。
542	多卡训练时常见的并行策略包括（ ）。
543	环境隔离工具的常见能力包括（ ）。
544	GPU 推理环境准备时常包含的步骤包括（ ）。
545	显存 OOM 时的常见排查思路包括（ ）。
546	评估报告常包含的内容包括（ ）。
547	确保实验可复现的常见做法包括（ ）。
548	常见的自动评估指标包括（ ）。
549	提交结果时通常需要附上的内容包括（ ）。
550	评估指标设计时常考虑的维度包括（ ）。
551	实验设计中常见的对照设置包括（ ）。
552	人工评估常见的做法包括（ ）。
553	评估报告中常见的数据可视化方式包括（ ）。
554	提交最终结果前常见的自检步骤包括（ ）。
555	评估多个模型时常用的统计检验方法包括（ ）。
556	LoRA 微调训练中影响最终效果的关键因素包括（ ）。
557	QLoRA 训练的关键组件包括（ ）。
558	完整 RAG 系统常见的关键模块包括（ ）。
559	向量检索系统常见的核心组件包括（ ）。
560	Transformers 推理服务的常见性能优化手段包括（ ）。
561	影响大模型推理时延的常见因素包括（ ）。
562	客服数据预处理的标准环节包括（ ）。
563	数据集融合过程中需要对齐的关键维度包括（ ）。
564	Prompt 工程中常用的核心技巧包括（ ）。
565	高质量 Prompt 设计需要考虑的因素包括（ ）。
566	数据探索阶段的常用分析维度包括（ ）。
567	监督微调数据预处理的关键步骤包括（ ）。
568	模型推理效果对比中常见的评测维度包括（ ）。
569	推理参数中影响生成结果的包括（ ）。
570	模型量化的常见目标包括（ ）。
571	选择模型部署方案时常考虑的因素包括（ ）。
572	GPU 推理环境常见的依赖项包括（ ）。
573	显存优化的常见手段包括（ ）。
574	完整实验报告通常包含的章节包括（ ）。
575	保证实验可复现的关键要素包括（ ）。
576	LoRA 微调训练中常见的数据准备步骤包括（ ）。

577	LoRA 训练中常见的超参数配置包括（ ）。
578	QLoRA 训练中通常需要关注的显存相关组件包括（ ）。
579	LoRA 训练完毕后的发布形式常见的有（ ）。
580	PEFT 库支持的低秩微调方法包括（ ）。
581	LoRA 微调相比全参数微调的典型优势包括（ ）。
582	LoRA 训练过程监控的常见指标包括（ ）。
583	LoRA 推理部署时常见的加载方式包括（ ）。
584	RAG 系统常见的检索优化策略包括（ ）。
585	向量数据库的常见评估维度包括（ ）。
586	RAG 系统中常见的文本清洗操作包括（ ）。
587	RAG 系统的 Prompt 拼接中常见的字段包括（ ）。
588	提升 RAG 答案质量的方法包括（ ）。
589	RAG 系统的常见工程指标包括（ ）。
590	构建多模态 RAG 系统常见的索引对象包括（ ）。
591	RAG 中对长文档的常见切分策略包括（ ）。
592	影响 Transformers 推理吞吐的常见因素包括（ ）。
593	KV cache 的工程优化方向包括（ ）。
594	高并发推理服务常见的可观测性指标包括（ ）。
595	推理服务中常见的限流策略包括（ ）。
596	Flash Attention 的工程收益包括（ ）。
597	Transformers 推理服务常见的部署形态包括（ ）。
598	推理引擎常见的功能包括（ ）。
599	Transformer 长上下文推理面临的挑战包括（ ）。
600	客服数据质量评估常用的维度包括（ ）。
601	客服对话样本中需要过滤的典型问题包括（ ）。
602	数据集融合时常见的对齐工作包括（ ）。
603	客服对话数据的常见标注字段包括（ ）。
604	客服样本数据增强常见的手段包括（ ）。
605	客服多轮对话数据常见的存储格式包括（ ）。
606	客服数据样本均衡的常见方法包括（ ）。
607	客服对话数据隐私脱敏的常见处理方式包括（ ）。
608	Prompt 工程中常用的引导技巧包括（ ）。
609	Few-shot 示例选择中通常关注的属性包括（ ）。
610	Chain-of-Thought 的典型使用场景包括（ ）。
611	ReAct 框架的典型组成包括（ ）。
612	System Prompt 设计中常见的约束包括（ ）。
613	提升 Prompt 鲁棒性的常见方法包括（ ）。
614	Prompt 评估的常见维度包括（ ）。
615	Instruction Template 中常见的占位字段包括（ ）。
616	微调数据预处理常见的步骤包括（ ）。
617	数据探索阶段常用的可视化包括（ ）。
618	数据切分时常用的策略包括（ ）。

微调数据预处理的 Token 化阶段常见的设置包括（）。

数据质量评估的常见维度包括（）。

数据集中常见的不良样本类型包括（）。

Instruction 模板拼接的常见字段包括（）。

样本标注一致性检验方法包括（）。

推理效果对比实验中常见的指标包括（）。

采样参数中影响多样性的包括（）。

提升推理效果的常见做法包括（）。

A/B 测试中常见的实验设计要素包括（）。

推理服务常见的性能优化方向包括（）。

对比实验中保证公平性的要素包括（）。

评估生成效果时常用的人工维度包括（）。

减少幻觉问题的常见方法包括（）。

模型量化的常见目标包括（）。

常见的量化方法包括（）。

GGUF 格式常见支持的特性包括（）。

选择模型部署方案时常考虑的因素包括（）。

ONNX 部署常见的优势包括（）。

TensorRT 部署中常见的优化包括（）。

4-bit 量化常见的精度损失来源包括（）。

部署格式选型时常考虑的生态因素包括（）。

GPU 推理环境常见的依赖包括（）。

显存优化的常见手段包括（）。

环境依赖管理常见工具包括（）。

CUDA 环境检查常见的命令包括（）。

多卡训练常见的并行策略包括（）。

模型加载时显存优化的常见手段包括（）。

环境复现常见做法包括（）。

显存 OOM 时的常见排查方向包括（）。

评估报告常见章节包括（）。

保证实验可复现的关键要素包括（）。

结果提交时常见的附件包括（）。

自动评估指标的常见类别包括（）。

实验设计中常见的对照设置包括（）。

人工评估常见做法包括（）。

评估报告常见的数据可视化包括（）。

提交前的常见自检项包括（）。

LoRA 通过冻结基座模型参数并仅训练低秩矩阵来减少显存占用。

QLoRA 在 LoRA 的基础上引入了 4-bit 量化以进一步降低显存。

LoRA 中 lora_alpha 越大，低秩更新的幅度通常越大。

PEFT 库中可以使用 merge_and_unload 将 LoRA 权重合并到基座模型。

LoRA 中 rank 越大，显存与过拟合风险都更高。

QLoRA 的 NF4 数据类型是按正态分布设计的非均匀量化。

QLoRA 训练中常用 gradient checkpointing 以进一步降低显存。
LoRA 训练 loss 出现 NaN 通常与学习率或异常输入有关。
DoRA 是一种 LoRA 变体,将权重分解为幅度与方向分别更新。
多任务场景下,使用独立 LoRA adapter 并动态加载通常能减少任务间干扰。
在 70B 模型上做 QLoRA 微调时,加载基座阶段最容易出现 OOM。
RAG 通过检索外部知识来补充模型回答,可以减少幻觉。
RAG 中向量数据库用于高效存储与检索文本向量。
RAG 中 top_k 越大,检索到的片段越多,但可能引入噪声。
RAG 中可以使用 BM25 与向量检索做混合检索。
Rerank 可以提升检索结果的相关性,通常放在向量初筛之后。
RAG 中切片粒度过大会导致噪声多且召回精度下降。
RAG 中 embedding 模型换为更强的通常可以提升召回质量。
RAG 中召回质量主要取决于 embedding 模型与切片粒度。
RAG 系统的延迟主要来自检索与 LLM 生成。
RAG 中对含表格的文档,按结构化方式切片通常更有利于保留信息。
GraphRAG 用图结构刻画实体关系,适合需要跨片段推理的场景。
RAG 评估的检索质量指标包括 Recall@k、MRR、nDCG。
HyDE 通过让模型先产生假设性文档再 embedding,可以缓解 query-doc 语义不对称。
transformers 中 model.generate 支持批量推理。
KV cache 通过缓存已计算的 Key/Value 来加速自回归生成。
attention_mask 用于标记 padding 位置以避免其参与注意力计算。
Python 的 GIL 不会影响多线程在 IO 密集型任务上的加速效果。
Flash Attention 通过分块与重计算降低显存并加速 attention。
use_cache=False 会让生成速度变慢。
vLLM 通过 PagedAttention 高效管理 KV cache 来提升吞吐。
对 decoder-only 模型做批量生成时,padding_side='left' 通常更合适。
推理时 GPU 利用率长期偏低通常是 batch 过小。
Speculative Decoding 通过小模型草拟、大模型验证来降低有效步数。
PagedAttention 主要解决长序列 KV cache 显存碎片问题。
Prefix Caching 可以让多轮对话中相同前缀共享 KV cache。
对超长上下文推理,KV cache 占用随序列长度线性增长。
Continuous Batching 在多请求混部下能显著提升吞吐。
torch.cuda.is_available() 可以判断当前是否能调用 GPU。
nvidia-smi 可以查看当前 GPU 显存占用。
加载大模型时设置 torch_dtype=torch.float16 可以降低显存占用。
出现 CUDA OOM 时减小 batch size 通常能直接缓解。
CUDA 与 PyTorch 版本不匹配可能导致 ImportError 或运行时错误。
low_cpu_mem_usage=True 可以减少大模型加载时的 CPU 内存峰值。
device_map='auto' 由 accelerate 自动决定模型各层所在设备。
「Expected all tensors to be on the same device」通常是因为部分张量未 .to(dev
safetensors 相比 pickle-based bin 在加载时更安全。
NCCL 主要针对 GPU 多卡高带宽通信优化。
对超大规模模型加载,使用 CPU offload 与 NVMe offload 可以避免单点 OOM。
环境自检通常需要采集 Python/PyTorch/CUDA 版本与 GPU 信息。
异构 GPU 集群部署推理时,通常需要针对不同型号做兼容与调度策略调整。
客服对话数据中敏感字段应进行脱敏后再使用。
客服对话常见存储格式包括 JSON、CSV、数据库 dump。
客服对话中电话号码通常应脱敏为 138****0000 形式。
客服数据多源融合时,首要工作是统一 schema 字段命名。
客服数据中可使用 sentence embedding + 余弦相似度检测语义重复。
客服多轮对话通常按窗口切片并构造 (instruction, response) 对进行训练。

客服数据评估集与训练集按时间或用户维度切分可以减少信息泄漏。
客服对话中重复问题会放大某些模式,建议去重或合并。
客服数据集中标签噪声过高会显著影响模型效果上限。
客服数据集可使用正则与 NER 模型识别 PII 并脱敏。
SimHash 比精确字符串匹配更适合客服对话的近似去重。
客服多任务标签(意图+情绪+槽位)通常应统一 schema 并独立标注。
客服对话与外部百科数据融合时,应建立权威优先级并记录证据来源以消解冲突。
GPTQ 是一种训练后量化方法。
4-bit 量化相比 FP16 会进一步降低显存占用。
bitsandbytes 中可以通过 load_in_4bit=True 加载 4-bit 模型。
ONNX 是一种跨框架的模型表示格式。
GGUF 主要面向本地 CPU/GPU 推理部署。
TensorRT 主要用于 GPU 上的推理加速。
动态量化与静态量化的关键区别在于 scale 计算时机。
SmoothQuant 主要解决激活值离群点导致的量化困难。
AWQ 基于激活分布识别并保护显著权重通道。
量化过程中校准数据不具代表性会导致精度掉点严重。
LoRA 在训练时会把基座模型的全部权重也一起更新以提升效果。
QLoRA 通过把基座量化为 FP8 来达到极致省显存。
LoRA 中 lora_alpha 与 rank 越大,模型一定越容易欠拟合。
PEFT 库对 LoRA 的支持与是否合并权重无关,合并仅用于 vLLM 推理。
LoRA 训练完成后,基座模型权重与原始版本已经完全不同。
QLoRA 的 NF4 数据类型会让精度与 FP32 完全一致。
LoRA 微调后无需任何处理即可直接对接到 TensorRT-LLM。
在 70B 模型上做 QLoRA 微调通常比在 7B 模型上节省更多显存。
RAG 完全消除了大模型的所有幻觉,所以无需做事实核验。
RAG 的向量数据库一定只存原文而不存元信息。
RAG 中 top_k 越大,生成答案一定越准确。
RAG 只能用向量检索,不能融合 BM25 等稀疏检索。
RAG 中切得越细召回越好,所以应尽量切到一句话一个 chunk。
RAG 检索阶段与生成阶段完全可以并行执行而不影响结果。
GraphRAG 在所有 RAG 场景中都优于朴素向量 RAG。
HyDE 通过让检索模型先生成假设文档,完美解决了所有 query-doc 不匹配问题。
transformers 中 model.generate 不支持多请求并发,只能一条条串行跑。
KV cache 在所有场景下开启都能加速,不必考虑显存。
Flash Attention 仅在 A100 上可用,其他 GPU 完全不兼容。
vLLM 只能跑在单卡上,多卡部署需要完全重写服务代码。
decoder-only 模型批量推理时,padding_side 必须固定为 right 才能保证正确性。
Speculative Decoding 通过大模型草拟、小模型验证来获得加速。
Continuous Batching 只在小 batch 场景下有收益,大 batch 下基本无效。
torch.cuda.is_available() 返回 True 就一定能完整跑通大模型训练。
nvidia-smi 不支持查看显存占用,只能用 torch 内部 API。
加载大模型时 torch dtype=torch.float64 比 float16 更省显存。
CUDA OOM 时增大 batch size 通常能直接缓解。
low_cpu_mem_usage=True 在加载完成后仍会显著增加 CPU 内存占用。
device_map='auto' 会无视硬件情况把整层模型放在 CPU。
异构 GPU 集群上所有推理框架都自动支持,无需任何适配。
客服数据脱敏只能由人工完成,模型无法辅助识别 PII。
客服对话中所有表情符号都应在清洗阶段全部剔除。
客服数据多源融合可以直接拼接,无需统一字段命名。
客服对话的寒暄部分一律删除即可,不会带来任何信息损失。

客服数据与通用指令数据融合训练一定能让客服效果提升。
客服多轮对话中,只标注最后一轮即可让模型学会上下文。
客服评估集与训练集随机打乱切分就足以避免信息泄漏。
客服对话与外部百科数据融合时,无须处理来源冲突。
GPTQ 必须在训练阶段对模型一起量化,不能在已训练好的模型上做。
4-bit 量化一定比 FP16 在所有任务上都更好,无任何精度损失。
bitsandbytes 只能支持 8-bit 量化,不支持 4-bit。
ONNX 仅能在 PyTorch 中生成,无法从其它框架导出。
GGUF 只能跑在 GPU 上,CPU 推理完全不可用。
动态量化与静态量化在精度和性能上完全一致。
QLoRA 微调后必须把模型反量化为 FP16/FP32,4-bit 无法直接部署。
temperature=0 时模型一定没有任何随机性,也不会因种子不同而变化。
top_p 越接近 1,生成越确定、越不容易重复。
repetition_penalty 设得越大,生成结果一定越好。
事实型问答应使用高 temperature 才能拿到最好效果。
A/B 测试中只要最终某组胜出即可,流量分布不用控制。
客服场景使用 temperature=0 通常更稳。
对比推理引擎吞吐时 P99 比 P50 更能反映典型用户体验。
SFT 不需要任何带标签数据,纯文本续写即可。
tokenizer 在不同模型间是通用的,可以随意替换。
padding 只会增加 batch 内最长样本的额外开销,不增加显存。
truncation=True 会让超长样本被随机丢掉,不会出现截断。
SFT 中对 prompt 和 response 一起算 loss 与只对 response 算 loss 完全等价。
SFT 训练中 loss 持续不下降通常说明数据规模足够,无需调整学习率。
SFT 数据中 prompt 与 response 的最佳比例是 9:1,资源越多越应偏向 prompt。
Few-shot 在 prompt 里塞越多示例就一定越好,不存在边际收益递减。
Chain-of-Thought 只能用于数学题,无法迁移到其他任务。
System Prompt 只对首次回答生效,多轮对话之后会失效。
ReAct 框架中 Reasoning 与 Acting 必须严格交替,不可并行。
JSON 严格输出无需给出 schema,只要在指令里说「请输出 JSON」即可。
多轮对话中,即便上下文已超长也应保留全部历史,删除会丢失关键信息。
Prompt Injection 攻击只在英文场景下有效,对中文 prompt 完全免疫。
BLEU 越高就代表生成质量一定越好,完全无需人工复核。
ROUGE 越高代表召回越好,与生成质量无关。
BERTScore 完全克服了 BLEU/ROUGE 的所有局限,适合所有任务。
客服比赛只看 BLEU/ROUGE 就足以评价综合效果。
比赛中只提交最终结果文件即可,无需写任何文档说明。

题型（必填）	选项A（必填）
单选	torch.cuda.is_available()
单选	只执行 del model 即可立即释放
单选	降低加载瞬间的 CPU 内存峰值
单选	df.isnull().sum()
单选	0
单选	重新训练全部模型参数
单选	使用 4-bit 量化加载基座以降低显存
单选	两组使用完全不同的测试问题
单选	PeftModel.from_pretrained(基座, 适配器路径)
单选	同时常驻两个模型于显存
单选	提升模型训练速度
单选	使点积结果即为余弦相似度
单选	生成文本与参考文本的 n-gram 重叠程度
单选	评测集与训练集尽量不重叠
单选	不提供示例,直接给出任务与输入
单选	在提示中明确 JSON schema 并给出示例
单选	正则表达式匹配对话标记
单选	按问题文本的哈希值去重
单选	设定问题与回答的最小/最大长度阈值
单选	串行逐条推理
单选	左填充(left padding)
单选	大幅降低显存占用

单选	对量化常数再做一次量化,进一步节省显存
单选	根据硬件约束与实际业务需求综合权衡
多选	Python 与 PyTorch 版本
多选	模型路径应可通过配置或环境变量统一修改
多选	统计样本数量与字段名称
多选	应区分 NaN 与空字符串
多选	对文本进行分词
多选	使用 4-bit 量化加载基座
多选	完全等价于直接使用更大的 batch_size
多选	固定随机种子保证可复现
多选	测试集覆盖多个业务场景
多选	只凭单条样例即可下结论
多选	每个字符单独作为一个 chunk
多选	随机梯度下降优化器
多选	返回片段与查询的相关性
多选	BLEU
多选	指标汇总表
多选	零样本提示
多选	提供与任务格式一致的示例
多选	引导模型分步骤推理
多选	正常的中文句子
多选	文件名拼写是否正确
多选	吞吐量(QPS)
多选	线程并发能让单卡计算真正并行
多选	控制 batch_size 不宜过大

多选	int64 高精度
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
单选	r
单选	bitsandbytes
单选	load_lora_model
单选	model.fuse_lora()
单选	lora_scale
单选	bf16/fp16
单选	embedding, lm_head
单选	hidden_dropout
单选	model.dump_weights()
单选	降低学习率并适当增 warmup_ratio
单选	FP4 精度高于 NF4
单选	pipeline parallel
单选	需要彻底改变模型架构
单选	r 越大可表达容量越大,但显存与过拟合风险也增加
单选	训练 loss 绝对值
单选	两者完全等价
单选	扩大 batch size
单选	merge and unload 合并到基座权重
单选	后者覆盖更广
单选	无需任何操作
单选	关闭梯度累积
单选	是否过拟合,考虑减小 rank 或增加 dropout
单选	提升训练精度
单选	全参数微调参数更少

单选	LoRA 减少了模型层数
单选	每个任务训练独立 adapter,推理时按任务切换
单选	保存 checkpoint 时
单选	完全替换 attention
单选	全部合并后只保留一份
单选	检索增强生成
单选	判别器
单选	压缩模型权重
单选	替代 embedding
单选	MySQL
单选	生成的最大长度
单选	Jaccard 系数
单选	BERT-base-uncased
单选	以独立 Context 段落形式置于 system 或 user 消息中
单选	增加 top_k 到 1000
单选	向量维度上升
单选	扩展上下文
单选	Hybrid Search 关键词+向量混合检索
单选	保持不变
单选	关闭向量检索改用纯生成
单选	随机切分
单选	embedding 模型与文档切片的语义粒度
单选	整篇合并
单选	只保留第一段
单选	操作系统调度
单选	扁平 brute-force 索引(规模小)
单选	完全替代正文
单选	磁盘带宽
单选	检索速度必然提升
单选	Recall@k、MRR、nDCG
单选	直接调小模型
单选	完全不需要向量库
单选	Python 版本不一致
单选	用图结构刻画实体关系,适合需要跨片段推理的场景
单选	减少 embedding 计算量
单选	model.forward(batch)
单选	蒸馏
单选	缓存已计算的 Key/Value,避免重复计算以加速自回归生
单选	控制学习率
单选	流式数据预处理
单选	Linux 内核
单选	使用 multiprocessing 或 Rust 实现的 fast tokenizer
单选	模型参数量
单选	关闭 KV cache
单选	字符串拼接
单选	通过分块与重计算降低显存并加速 attention

单选	训练速度下降
单选	逐条串行推理
单选	无需 GPU
单选	推理阶段的超参数(温度、top_p、max_new_tokens 等)
单选	使用统一 batch size
单选	CPU 解码
单选	embedding 检索
单选	提高 batch size 与连续批处理(continuous batching)
单选	padding 标记
单选	与图像相同
单选	降低 CPU 占用
单选	Continuous batching(连续批处理)
单选	增加 batch size
单选	DeepSpeed ZeRO
单选	指数增长
单选	通过算子融合与图优化减少 kernel 启动开销
单选	广播 embedding
单选	关闭 attention
单选	torch.device('gpu')
单选	nvidia-smi
单选	Model.load
单选	兼容旧 GPU
单选	关闭 SSD
单选	accelerate
单选	随机分配权重
单选	pt.ver
单选	训练 epoch 翻倍
单选	使用 FP16/BF16 半精度加载
单选	立即换更大显卡
单选	更换浏览器
单选	CPU 不支持
单选	FP16 + KV cache 共享 + 短上下文
单选	ls transformers
单选	重启 GPU
单选	Python 路径错误
单选	加载时量化(4-bit/8-bit)或上 ZeRO/CPU offload
单选	训练数据
单选	加快 GPU 推理
单选	安装额外 GPU
单选	Python/PyTorch/CUDA 版本、GPU 型号、显存、是否可用
单选	全部跨机同步
单选	CUDA 未安装
单选	随机初始化加载
单选	accelerate
单选	完全随机
单选	网络问题
单选	Gloo 不支持分布式

单选	分层 offload 到 CPU 与 NVMe,并流式加载
单选	训练模型
单选	直接删除文件
单选	MP3
单选	按对话指纹(如 MD5/SimHash)或语义相似度去重
单选	保留原值
单选	直接拼接
单选	英文对话
单选	GPU 型号
单选	A / B / C
单选	数据丢失
单选	比较字数
单选	过滤或下采样营销类样本
单选	随机打散
单选	改文件名为 hash
单选	随机排序
单选	评估其信息量后选择保留或规范化映射
单选	推理速度下降
单选	丢弃 FAQ
单选	减少显存
单选	最近若干轮 + 系统指令 + 用户最新问题
单选	完全随机洗牌
单选	全部删除文件
单选	按字数截断
单选	对近似文本敏感且可大规模比对
单选	增加数据量

单选	明文存储
单选	只标意图
单选	按主题聚类并裁剪跨主题的拼接样本
单选	直接丢弃附件信息
单选	全部保留原文
单选	改变 attention
单选	训练后量化方法
单选	显存占用更高
单选	use_4bit=True
单选	AWQ 仅支持 CPU
单选	跨框架部署与跨硬件加速
单选	训练加速
单选	数据标注
单选	不量化但乱用设备
单选	GGUF + llama.cpp
单选	必须反量化到 FP32
单选	LlamaForCausalLM
单选	训练轮数太多
单选	动态量化在推理时计算 scale,静态量化用校准集预先计
单选	GPU 不够
单选	INT2
单选	二值化
单选	检查量化校准数据分布与异常输入
单选	权重更大
单选	FP32 加载
单选	Python 版本
单选	在 bitsandbytes 下继续 4-bit 推理,显存最低
单选	完全不影响精度
单选	只支持图像
单选	校验输出
单选	显存减半,部分硬件吞吐翻倍
单选	纯 CPU + FP32
单选	保护全部权重
单选	CPU 完整加载
单选	temperature
单选	[1, ∞)
单选	完全随机
单选	仅取最后一个 token
单选	惩罚重复生成以提升多样性
单选	高 temperature + top_k=1
单选	只有 grad
单选	tokenizer 改变
单选	两组流量分布与请求特征基本一致
单选	极低 temperature + greedy
单选	加大 batch
单选	只看 BLEU

单选	逐个参数网格搜索并记录指标
单选	每轮重置上下文
单选	token 总数
单选	beam search 大小为 1 但 temperature=2
单选	temperature 升高 + top_p 升高 + repetition_penalty 降低
单选	只看 BLEU
单选	A 用 system B 不用
单选	TopKConfig
单选	repetition_penalty 过低或未设置
单选	增加 temperature
单选	固定采样
单选	GPU 不一致
单选	灰度放量 + 监控关键业务指标 + 异常回滚预案
单选	纯 greedy
单选	temperature=2
单选	只展示 GPU 利用率
单选	相同硬件/模型/请求分布下多轮压测取 P50/P99
单选	加快 GPU
单选	无监督聚类
单选	(audio, text)
单选	将文本转为模型可处理的 token id
单选	SFT 数据量更大
单选	增加 token
单选	删除停用词
单选	只计算 response 部分的 loss,屏蔽 prompt
单选	压缩权重
单选	雷达图
单选	加大 batch
单选	按窗口切片或丢弃过长样本
单选	按文件名
单选	显存暴涨
单选	只看显存
单选	训练指标虚高,推理效果差

单选	全部为 1
单选	提升 GPU 性能
单选	样本 ID 奇偶
单选	风格单一,泛化能力差
单选	全部保留
单选	随机打散无标识
单选	随机
单选	loss masking 错误,模型学到了 prompt 而非 response
单选	随机切分
单选	替换 tokenizer
单选	改 batch size
单选	评估模型泛化与鲁棒性
单选	加大 batch
单选	删除历史
单选	删除全部
单选	在 prompt 中给出若干示例引导模型
单选	提供 5 个示例
单选	随机回答
单选	增加 batch
单选	Reasoning + Acting
单选	只给用户问题
单选	tokenizer
单选	无任何约束
单选	在 prompt 中给出具体要求与格式示例
单选	加速推理
单选	随机回答
单选	替换 attention
单选	明确 Context + Instruction + Question + 引用要求
单选	同时改 5 个变量
单选	显存下降
单选	禁用回答
单选	展示推理过程便于审计与排错
单选	压缩权重
单选	强制截断
单选	只保留最后一句
单选	结构化、分点、加入示例与边界条件

单选	降低 temperature
单选	显存不足
单选	添加越权检测
单选	思考(Thought)与行动(Action)交替
单选	随便描述
单选	加入示例
单选	更换模型
单选	Plan-Then-Execute + 显式校验与回退
单选	图像质量
单选	显存
单选	更快
单选	CSV/JSON 含指定字段(如 id, answer)
单选	只看 BLEU
单选	关闭日志
单选	GPU 型号
单选	多指标组合(自动 + 人工)综合判断
单选	指标更准
单选	只看字符
单选	隐藏指标
单选	提交前固定 shuffle 种子或显式排序
单选	只看一次
单选	显卡型号
单选	只给一行结论
单选	本地评测集分布与线上不同,或线上规则更严
单选	凑字数
单选	散点图
单选	错误率
单选	附带可复现推理脚本与依赖版本
单选	只固定一处
单选	只看人工评分
单选	关闭 prompt
单选	客观陈述已知不足并给出后续改进思路
单选	只看主流集
单选	在线下载
单选	只看最大值
单选	建立红线 prompt 与内容过滤并记录审计日志
单选	只看一次
单选	不使用
多选	W + A·B
多选	NF4 数据类型
多选	更新幅度缩放
多选	AutoModel.from_pretrained
多选	验证指标
多选	Embedding 层
多选	load_adapter(adapter_name)
多选	通常设为 0.05–0.1
多选	学习率震荡
多选	数据量减少
多选	固定长度切分

多选	Top-K 过滤
多选	BLEU 分数
多选	写入吞吐
多选	加快推理速度
多选	模型微调
多选	稠密向量检索
多选	生成最终回答
多选	模型超参数
多选	训练 Loss
多选	KV cache
多选	vLLM
多选	替换模型架构
多选	解码策略
多选	替换模型权重
多选	训练集大小
多选	模型大小
多选	显著降低显存
多选	Embedding 拼接
多选	关闭 KV cache
多选	表情符号与乱码
多选	问题分类标签
多选	随机哈希
多选	提升 Batch Size
多选	调高学习率
多选	模型架构
多选	直接删除
多选	模型结构一致
多选	高质量标注
多选	模型 Loss
多选	示例数量适中
多选	让模型只输出关键词
多选	反向传播
多选	设定训练 Loss
多选	学习率
多选	增加模型层数
多选	###
多选	示例越多越好
多选	显著降低显存
多选	忽略所有指令
多选	prompt
多选	训练时统一截断到 4K
多选	测试集必须大于训练集

多选	模型 Loss
多选	随机打乱标签
多选	样本长度统计
多选	长度分布直方图
多选	随机重命名
多选	随机丢弃字符
多选	格式统一标准样本
多选	控制采样随机性
多选	替换模型架构
多选	训练轮次
多选	提升 Batch Size
多选	模型参数量差异
多选	数据增强
多选	只跑一次
多选	磁盘容量
多选	多次采样取平均
多选	显存占用
多选	GGUF
多选	Loss 函数改变
多选	训练数据量
多选	只能用于训练
多选	随机化权重
多选	自带训练能力
多选	算子融合
多选	Loss vs 精度
多选	提高 batch_size
多选	训练日志清晰度
多选	关闭无用进程
多选	ls /dev/null
多选	txt 笔记
多选	调大 Batch Size
多选	把模型完整加载到 CPU 再一次性搬上 GPU
多选	训练集大小
多选	DDP
多选	是否支持图像编辑
多选	安装 Photoshop
多选	调高 epoch
多选	实验目的
多选	每次都换数据集
多选	显存峰值
多选	显卡 BIOS
多选	是否好看

多选	无对照组
多选	多人盲评
多选	Photoshop 海报
多选	重命名训练集
多选	直接肉眼对比
多选	Attention 的 q_proj
多选	4-bit 量化基座
多选	lora_rank
多选	训练完毕上线
多选	save_pretrained 保存
多选	更新幅度缩放
多选	显存占用更低
多选	LoRA
多选	EarlyStopping
多选	多个 adapter 命名隔离
多选	固定长度滑窗切分
多选	余弦相似度
多选	提高 Top 结果相关性
多选	是否好看
多选	答案准确性
多选	文档预处理
多选	限制上下文长度
多选	暴力搜索
多选	保留重叠窗口
多选	Hit Rate
多选	复用历史 token 的 K/V
多选	vLLM
多选	降低显存
多选	请求按到达时间入队
多选	PagedAttention
多选	temperature
多选	减少计算复杂度
多选	模型加载太慢
多选	Multi-Query Attention
多选	temperature
多选	去重
多选	手机号
多选	完全重复
多选	同义改写
多选	过采样少数类
多选	ShareGPT 风格
多选	用户骂人

多选	字段重命名
多选	回答专业性
多选	正则替换
多选	示例数量
多选	Zero-shot CoT
多选	Thought
多选	回答长度限制
多选	Instruction
多选	A/B 测试
多选	明确任务说明
多选	示例尽可能多
多选	Let’ s think step by step
多选	使用固定 token 包裹
多选	Alpaca
多选	统一 batch 内序列长度
多选	独立同分布
多选	覆盖率
多选	使用统一模板
多选	双人标注
多选	样本总数
多选	字段重命名
多选	原始数据清洗
多选	超长样本截断
多选	T=0 接近贪心
多选	限制候选 token 数量
多选	BLEU
多选	模型不变只改 prompt
多选	流量分桶
多选	KV cache 复用
多选	固定测试集
多选	磁盘容量
多选	准确性
多选	加入 RAG 检索
多选	权重舍入误差
多选	单文件部署
多选	量化算法不同
多选	训练 Loss 大小
多选	跨平台
多选	算子融合
多选	GGUF
多选	训练集大小
多选	困惑度对比

多选	加载校准集
多选	关闭无用进程
多选	nvidia-smi
多选	pip
多选	调大 Batch Size
多选	低精度加载
多选	torch 版本
多选	DDP
多选	不支持 GPU
多选	安装 NVIDIA 驱动
多选	查看 nvidia-smi
多选	实验背景
多选	固定随机种子
多选	BLEU
多选	显卡 BIOS
多选	有效性
多选	基线模型
多选	多人盲评
多选	Photoshop 海报
多选	核对格式
多选	配对 t 检验
多选	数据集质量
多选	4-bit 量化基座
多选	文档切片模块
多选	Embedding 模型
多选	KV cache 复用
多选	模型参数量
多选	去重
多选	Schema 字段
多选	Few-shot 示例
多选	明确的指令
多选	样本总量
多选	数据清洗
多选	生成质量
多选	temperature
多选	降低显存
多选	硬件兼容性
多选	NVIDIA 驱动
多选	混合精度
多选	实验背景
多选	随机种子固定
多选	原始语料收集

多选	lora_rank
多选	4-bit 基座
多选	单独 adapter 文件
多选	LoRA
多选	显存更省
多选	训练 Loss
多选	单独加载 adapter
多选	查询改写
多选	召回率
多选	去除 HTML 标签
多选	任务说明
多选	使用更好的 Embedding 模型
多选	检索时延
多选	文本段落
多选	固定窗口切分
多选	批处理大小
多选	压缩量化
多选	QPS
多选	QPS 限流
多选	减少显存
多选	单机单卡
多选	动态批处理
多选	KV cache 显存爆炸
多选	回答准确性
多选	答非所问
多选	字段重命名
多选	意图分类
多选	同义改写
多选	JSON Lines
多选	过采样
多选	正则替换
多选	明确角色
多选	多样性
多选	数学推理
多选	Thought
多选	回答长度
多选	显式边界说明
多选	任务完成率
多选	Instruction
多选	数据清洗
多选	长度直方图
多选	随机切分

多选	max_length
多选	覆盖率
多选	空回答
多选	任务说明
多选	双人标注
多选	首字时延
多选	temperature
多选	调参
多选	流量分桶
多选	KV cache 复用
多选	相同测试集
多选	准确性
多选	加入 RAG 检索
多选	降低显存
多选	PTQ（训练后量化）
多选	单文件部署
多选	硬件兼容性
多选	跨平台
多选	算子融合
多选	权重舍入误差
多选	推理引擎支持
多选	NVIDIA 驱动
多选	混合精度
多选	pip
多选	nvidia-smi
多选	DDP
多选	低精度加载
多选	固定依赖版本
多选	nvidia-smi 查看占用
多选	实验背景
多选	固定随机种子
多选	评估脚本
多选	BLEU
多选	基线模型
多选	多人盲评
多选	折线图
多选	核对字段
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确

判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确

判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确

判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确
判断	正确

选项B（必填）
torch.gpu.is_ready()
del 引用后配合 gc.collect() 与 torch.cuda.empty_cache()
减少模型的 GPU 显存占用
df.count()
1
冻结基座,只训练注入的低秩矩阵 A/B
使用更大的学习率
两组使用相同的测试问题与生成设置
直接替换模型权重文件
依次加载、用完即释放,避免峰值叠加
为生成补充外部知识、缓解幻觉
提高模型精度
模型推理速度
评测集越大越好且必须来自训练集
提供大量标注示例
完全依赖模型自由发挥
随机抽样
按回答长度去重
只保留最长的样本
多线程并发推理
右填充(right padding)
提升模型精度

提高模型输出精度
永远选精度最高方案
CUDA 是否可用及 GPU 名称
把路径写死在代码里最便于维护
检查字段类型与典型样本
所有缺失值一律填 0 最安全
对数据做加密存储
开启梯度检查点
batch=1 配合累积可模拟更大等效 batch
记录训练损失等监控信息
只挑选表现最好的样本展示
可从回答长度观察风格变化
按段落或标题切分
卷积神经网络
知识库文件的大小
ROUGE-L
可视化对比图
少样本提示
示例越多越好,不受上下文限制
无条件加速模型推理
表情符号
模型目录是否存在
平均延迟
可重叠 tokenize 与后处理
限制最大序列长度

fp16 半精度
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
k
accelerate
get_peft_model
model.unload_adapter()
lora_dropout
fp32
q_proj, v_proj
lora_alpha
model.save_adapter()
把秩 r 翻倍
NF4 是按正态分布非均匀量化,更适合权重
tensor parallel
有充足的多机多卡预算
r 越大显存占用一定减少
下游任务的指标提升幅度
Adapter 改造权重矩阵,LoRA 插入模块
提升训练速度
继续保留 adapter 不合并
前者覆盖全部 Attention 投影,可表达容量更大
直接覆盖旧权重
直接更换基座模型
降低 batch size
缓解 optimizer 状态在长序列下的显存尖峰
LoRA 等价于全参数微调

LoRA 改变了模型结构
把所有任务数据混合成一份
加载基座模型到 GPU 时
将权重分解为低秩与稀疏两部分
为每个任务合并出独立全量模型
递归注意力图
Embedding 模型
训练大模型
节省存储空间
FAISS
检索返回的最相似片段数量
欧氏距离取负
ResNet
替换用户问题
更换或微调 embedding 模型,并优化切片粒度
embedding 计算变慢但更准
替换 embedding 模型
仅 BM25
适当调大,给 rerank 留出排序空间
直接换更大的模型
按字符固定长度切
GPU 型号
按表格/章节结构切分并附带标题元数据
随机选一段
GPU 启动
倒排索引(IVF)
在窗口有限时保留核心语义
GPU 显存
语义表达一定更好
BLEU
先看检索结果,再看出处片段,最后看 prompt 拼接与生成参数
维度更低
embedding 维度变化
使用更小的 embedding 维度
让查询与文档在同一空间对齐,缓解 query-doc 语义不对
model.predict(batch)
归一化
压缩模型权重
标记 padding 位置避免其参与注意力计算
流式加载模型权重
GPU 驱动
Python for 循环
decoder-only 模型的批量生成位置对齐
减小 batch 到 1
Python 字典查询
等价于 KV cache

显存占用随序列长度线性增长
全部 pad 到最长
支持更多 embedding 模型
训练超参数
频繁申请/释放不同 shape 的张量
逐条复制
训练
降低 batch size
触发生成终止的结束符
(L, L) 对角矩阵
替代 attention
固定 batch
用小模型草拟 token,大模型批量验证,降低有效步数
DataParallel
常数
改用 CPU
all-reduce 输出投影
丢弃所有 KV
torch.gpu.is_ready()
lsmod
AutoModelForCausalLM.from_pretrained
提升模型精度
增加 batch size
pandas
由 accelerate 自动将模型按层分配到多 GPU/CPU
torch.cuda.version()
loss 升高
使用 FP64
检查是否有未释放的中间变量或未调用 del 与 torch.cuda.empty_cache()
改 Python 版本
模型已损坏
FP32 + 长上下文 + 多 batch
pip show transformers
删除模型权重
数据缺失
直接 FP16 加载
模型结构与超参数配置
提高精度
device_map='cuda'
CPU 频率、硬盘品牌、鼠标型号
尽量在节点内做张量并行,跨机仅做流水线并行
显存不足
FP32 加载
numpy
更均匀地在多卡间分配层,但可能牺牲连续性
Python 路径缺失
Gloo 比 NCCL 更快

一次性全量预加载
去除无效字符与空对话
原样存储
PDF
随机删除一半
脱敏为 138****0000 形式
随机抽样
所有对话
乱码字符
user / assistant / system
训练变慢
比较字符串相等
全量训练
按对话窗口切片并构造（instruction, response）对
直接删除文件
直接复制少数类 1000 倍不加评估
全部删除
模型效果上限下降,且难以快速定位原因
直接拼接纯文本
毫无影响
开头一字
按时间或用户维度切分,避免信息泄漏
全部保留
随机合并
完全无碰撞
过滤低质量样本并加权高质量样本

直接发布模型
全部用同一标签
全部保留
对附件做 OCR/ASR 摘要并与文本对齐
随机选一个
提升训练精度
训练方法
精度损失更大
quantize_4bit=True
AWQ 不量化
更换训练框架
GPU 上推理加速与部署
训练
4-bit 量化全模型
TensorRT
将 4-bit 基座与 LoRA 合并后做 4-bit 推理
AutoModel
GPU 数量过多
动态量化更慢
激活值离群点导致量化困难
FP64
INT2 量化
更换 CPU
图固定、算子固定,跨平台部署友好但灵活度下降
直接 FP16 加载
模型名字长度
立即转 FP32
长序列场景显著降低显存且精度损失可控
不支持 FP16
选择量化方案
显存翻倍
TensorRT-LLM + 量化 + 连续批处理
随机保护
FP32 加载
batch size
(0, 1]
只保留 1 个候选
保留全部 token
强制重复
低 temperature + 中等 top_p
只有 loss
显存下降
使用相同模型不同 prompt
中-高 temperature + 较高 top p
关闭 GPU
单条样本人工判断

一次性调全部参数
拼接历史对话并稳定使用同一推理参数
单条回复长度
高 temperature + 大 top_p
全部设最小值
构建贴近真实业务的人工评分集
A 用长 prompt B 用短 prompt
GenerationConfig
batch 太大
开启 KV cache + 量化 + 连续批处理
完全随机
A 模型更差
全量替换
中 temperature + 较高 top_p + 适度 repetition_penalty
do_sample=True + top_k=1
只贴一张图
只跑一次
抑制不必要复读同时保留多样性
自监督学习
(image, label)
压缩模型
SFT 使用带标签的少量数据继续微调
删除数据
扩展长度
屏蔽全部 loss
规范多轮对话角色与分隔符,提升模型对角色感知
饼图
更换 CPU
全部 padding 到最大
按 instruction 或完整样本 embedding 去重
推理变慢
只看 loss 曲线
训练报错

确认 padding 位置被 mask 掉
压缩权重
字符数
推理速度下降
过滤或合并到上下文样本中
全部混在一起
英文专用
学习率过低
按对话回合滑动窗口并保留最近上下文
压缩数据
删除全部
压缩权重
减小学习率或提前停止并检查过拟合
只对最后一句 tokenize
直接训练
训练大量数据
不提供示例直接让模型完成任务
压缩答案
压缩模型
Read + Apply
在 system 中明确角色身份与边界
训练模式
随便回答
调大 batch
超出上下文窗口且稀释指令
单次推理
加快推理
只给问题
只改动一个变量,其他条件一致
推理变慢
必须给一个答案
压缩回答
在思考与行动之间交替,便于调用外部 API
让模型自由输出
删除全部历史
继续堆文字

在 prompt 中要求逐步显式推理
GPU 故障
使用分隔符与校验
固定答案
清晰描述工具签名与触发条件
结构化分块
完全无推理
直接执行
机器翻译/生成的 n-gram 重叠度
图像分类
只看字符重叠
PDF
业务指标(如解决率)+ 人工评分
随意调参
模型名字长度
只看 1 个指标
离线集与真实分布不一致或指标脱节
只看长度
提交超出约束
随机顺序
多次重复并使用配对检验(t-test/bootstrap)
模型名字
禁用 JSON
本地一定更准
帮助定位系统边界并提出改进方向
饼图
P50/P99 延迟
只提交权重
对所有随机源(seed、numpy、cuda)固定
只看离线
随意拼接
只夸优点
构建长尾集并与主流集对比
调用远程 API
只看一次
关闭过滤
相同硬件下重复测量取 P50/P99
隐藏截止时间
A·B 独立训练
Double Quantization
有效学习率大小
Trainer.train
GPU 温度
最终分类头
set_adapter(adapter_name)
用于增大学习率
数据加载瓶颈
训练 Epoch 缩短
按句子/段落切分

随机丢弃
训练 Loss
训练精度
降低显存
增量索引
BM25 关键词检索
做模型量化
学习率
数据量
动态批处理
LoRA 训练
提升词向量维度
训练 epoch
提升训练 Loss
学习率
输入长度
减少 token 数量
Loss 计算
提高 batch_size
重复对话模板
时间戳
数据长度比
改 Loss
加更多 epoch
优化器类型
替换为占位符
评测指标一致
标注一致性高
Batch Size
示例与任务高度相关
强制使用 JSON 格式
梯度更新
控制 GPU 显存
Batch Size
改用更大词表
``
只放同一种类
替代 fine-tuning
完全信任用户输入
system
强制 padding 到全局最大值
随机切分

Batch Size
去除样本
字段缺失率
词云
整列清空
调大学习率
高质量标注样本
影响生成多样性
替代 attention
学习率
替换数据集
训练数据差异
特征工程
控制变量
CPU 频率
固定 seed
训练时长
GPTQ
优化器差异
词表大小
需要特定 GPU
扩大词表
需绑定 PyTorch
kernel 自动调优
训练时长 vs 推理速度
CPU offload
显存占用
使用 gradient checkpointing
pip list
Excel 维护
禁用 cache
直接 float32 加载
学习率
FSDP / ZeRO
是否能剪辑视频
安装视频编解码器
更换数据集
实验方法
更换训练框架
Loss 曲线
操作手册
是否新颖

单组实验
打分表设计
手绘插画
删掉测试集
主观打分
MLP 的 up_proj
NF4 数据类型
lora_alpha
准备数据集
load_adapter 加载
有效学习率
可在 24GB 显卡上微调 7B 模型
Prefix Tuning
SaveCheckpoint
base model 共享
按语义段落切分
欧氏距离
减少喂给 LLM 的噪声
中文支持能力
上下文相关性
向量索引
防止 prompt 注入
HNSW
按章节标题切分
MRR
加速自回归生成
TGI
提升速度
统一 padding
Continuous Batching
top_p
减少显存占用
显存不足
Grouped-Query Attention
top_p
去噪
身份证号
模板近似重复
回译
欠采样多数类
Alpaca 风格
空白对话

标签体系对齐
回答完整性
哈希替换
示例多样性
Few-shot CoT
Action
输出格式要求
Input
Log 分析
明确输出格式
示例多样化
示例推理过程
在对话中复述要求
ShareGPT
限制最大长度
不泄漏测试集
准确性
模板中区分 system/user/assistant
抽样复核
字段缺失率
标签对齐
分词编码
空回答过滤
T=1 默认分布
影响多样性
ROUGE
更换更大模型
对比关键指标
量化
控制变量
训练数据大小
流畅性
加强指令遵循
激活值量化
多种量化类型
校准数据使用方式不同
硬件兼容性
跨框架
kernel 自动调优
ONNX
吞吐量
下游任务评测

逐层量化
使用 gradient checkpointing
nvcc --version
conda
禁用 cache
device_map='auto'
transformers 版本
ZeRO
依赖版本管理
安装 CUDA Toolkit
降低 batch
实验方法
记录超参数
ROUGE
实验配置
可计算性
消融实验
随机给分
折线图
核对字段
Wilcoxon 符号秩检验
lora_rank 取值
NF4 数据类型
向量索引模块
向量索引结构
动态批处理
上下文长度
脱敏
标签体系
Chain-of-Thought
清晰的输入
字段缺失率
格式转换
首字时延
top_p
加速推理
推理速度
CUDA Toolkit
梯度累积
实验方法
数据版本一致
格式标准化

lora_alpha
LoRA adapter
合并后全权重
QLoRA
训练更快
学习率变化
合并后整模型加载
HyDE 假设文档
查询时延
去除多余空白
检索上下文
增加 Rerank
召回率
图片特征
按语义切分
并发请求数
分页管理
首字时延
并发数限流
提升速度
单机多卡
KV cache 管理
推理变慢
覆盖场景
空回答
标签体系统一
情绪标签
回译
CSV
欠采样
哈希替换
指定输出格式
代表性
逻辑推理
Action
回答格式
Few-shot 示例
格式符合度
Input
格式转换
类别饼图
分层切分

padding 策略
标注一致性
超长 prompt
用户输入
抽样复核
端到端时延
top_p
改 Prompt
对照设置
模型量化
控制变量
流畅性
加强指令遵循
加速推理
QAT（量化感知训练）
多种量化方案
推理速度
跨框架
kernel 自动调优
激活值量化
社区活跃度
CUDA Toolkit
梯度累积
conda
nvcc --version
ZeRO
device_map='auto'
记录 requirements
降低 batch size
实验方法
记录超参数
模型 checkpoint 链接
ROUGE
消融实验
打分表设计
柱状图
核对格式
错误
错误
错误
错误
错误
错误

错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误

错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误

错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误
错误

选项C
torch.cuda.check()
必须重启内核才能释放
加快模型推理速度
df.describe()
-100
删除模型的部分网络层
使用更深的网络结构
只比较回答长度
重新训练一遍模型
只用基座模型即可
降低模型参数量
减少文本长度
模型显存占用
评测集只需一条样本
必须使用思维链
只给一句模糊指令
人工逐条手抄
按样本序号去重
删除所有中文样本
真批量(batched)推理
不填充
加快训练收敛

增加模型参数量
永远选显存最小方案
模型训练集的样本数量
运行期不得联网下载模型
分析序列长度分布
缺失值可以完全忽略不处理
拼接提问、回答与结束标记
增大 batch_size 以减少迭代次数
能在不增加显存的前提下扩大等效 batch
保存最终适配器权重
为每条样本保留类别标签
可从回答稳定性角度评估
切得越碎越好
向量存储与相似度检索
相似度分数的高低
模型文件大小
原始评测结果
反向传播算法
示例数量控制在 2~5 个为宜
减少模型参数量
正常标点符号
每行是否为合法 JSON
模型参数数量
受 GIL 限制计算难以真并行
显存不足时自动降级或报错处理

int8 量化

n

deepspeed

inject_adapter

model.merge_and_unload()

lora_rank

int8

mlp.up, mlp.gate

lora_dropout

model.export_lora()

禁用 gradient checkpointing

NF4 仅支持 CPU

gradient checkpointing

基座模型为从头训练的小模型

r 越小效果一定越好

基座模型的 perplexity

LoRA 通过低秩分解改造权重矩阵,Adapter 在网络中插入新模块

替代 optimizer

先卸载基座模型

两者等价

unload 旧 adapter 后重新注入

立即加大 batch size

关闭 warmup

替换梯度计算

LoRA 假设权重更新具有低秩结构,因此用更少参数近似

全参数微调使用了更小的 batch
持续增大 rank
打印日志时
将权重分解为幅度与方向并分别更新
放弃 adapter 改用 prompt
重参数对齐门
解码器
高效存储与检索文本向量
加快训练速度
Milvus
embedding 维度
余弦相似度
YOLO
作为 temperature 参数
关闭 rerank
噪声多,易超出上下文窗口,且召回精度下降
压缩向量
仅 dense retrieval
设为 1
检查检索片段是否真的与问题相关,以及 prompt 是否清晰引用了片段
整篇文档不切
Python 版本
逐字拆分
在 prompt 中要求模型基于多片段对比并指出冲突
Python 解释器
HNSW
减少 embedding 维度
内存读写与图遍历
存储开销不变
Perplexity
重启服务
在 token 粒度匹配,长文档细粒度召回更准
GPU 故障
完全去掉 LLM
替代向量库
model.generate(batch)
量化
提升 embedding 维度
选择优化器
逐 token 流式输出生成结果
CUDA
纯单线程
训练 loss
在显存允许范围内尽量增大 batch 并启用 KV cache
整数加法
只支持 CPU

loss 上升
按长度分桶(bucketing)后再 padding 批量推理
自动训练
优化器状态
启用 gradient checkpointing
使用连续 batch + 预 pin 的输入张量
数据标注
关闭 GPU
attention mask 标记
(B, L) 中 1 表示有效 token
提升 embedding
严格 FIFO
关闭 KV cache
单进程多卡 model parallelism
对数增长
替换 attention
单卡读取
共享相同前缀的 KV cache 以减少重复计算
torch.cuda.check()
free -m
open(model_path)
降低显存占用并加速推理
更换 CPU
requests
关闭 GPU
torch.__version__
显存暴涨
关闭 GPU
重装系统
检查显存是否足够、版本是否匹配、是否混合精度不一致
GPU 已坏
FP64 + 单 batch
echo transformers
释放 PyTorch 缓存的未占用显存给系统
网络断开
使用 FP64
优化器状态
减少加载阶段 CPU 内存峰值,避免大模型加载 OOM
torch_dtype=torch.float16
显示器分辨率
关闭 NCCL
state_dict 键与模型结构不匹配
FP64 加载
pandas
全部放 CPU
CUDA 运行时与 PyTorch 不匹配
NCCL 仅支持 Windows

关闭 swap
向量化
进行脱敏处理后再入库
BMP
按文件名
删除文件
统一 schema 字段命名与时间格式
长对话
表情符号
True / False
数据分布偏移导致通用能力退化
比较文件大小
删除全部
删除所有上下文
用正则与 NER 模型识别并替换/掩码
全部丢弃
全部转大写
显存占用下降
将 FAQ 转成指令-回答对再合并
一定提升效果
中间一字
训练集挑难的
保留少量用于礼貌风格,其余过滤或归一化
删除工单 ID
支持图像
减少训练时间

在数据管线中先正则+NER 识别再脱敏,带审计日志
随机生成标签
随机丢弃
随机保留
建立权威优先级并记录证据来源
替换 tokenizer
数据增强方法
推理更慢
load_in_4bit=True
GPTQ 速度更快
提升训练速度
数据清洗
本地 CPU/GPU 推理部署
二值化
DeepSpeed
删除基座模型
AutoGPTQForCausalLM
Python 版本老
静态量化不支持 LLM
Python 性能
GPTQ/AWQ/FP16 等
随机量化
关闭 SSD
训练更快
AWQ/GPTQ 4-bit 或合并 LoRA + 4-bit 量化
训练集大小
必须先 GPTQ 再推理
训练更快
原生支持 KV cache 优化与 in-flight batching
转换脚本
完全没区别
GGUF + CPU 单线程
通过激活分布识别并保护显著权重通道
二值化
lr
[-1, 1]
在每步保留多个候选序列
随机 k 个 token
增加 batch
完全 greedy + 无限制 max_new_tokens
sequences / scores 等字段
loss 升高
让 A 组更准
temperature=0 + top k=1
加大 repetition_penalty 或提高 temperature/top_p
选自己偏好的模型

固定所有参数
调高 temperature 到 2
Self-BLEU / n-gram 多样性 / 文本相似度方差
完全随机
全部设最大值
只看 ROUGE
使用同一 prompt 与推理参数
BeamConfig
top_k 太小
使用更大模型
结合概率绝对阈值与 top-p,过滤低概率长尾
B 模型更慢
直接关闭旧模型
temperature=2
do_sample=False + num_beams>=1
只给一行结论
只看显存
压缩权重
监督微调
(video, caption)
量化
SFT 不需要 GPU
将不同长度样本补齐到统一长度以便批处理
添加 padding
增加 loss
替换 tokenizer
直方图 / 分布图
关闭 SSD
修改模型位置编码
按字符长度
模型容易学成复读 prompt,生成质量下降
只看训练时长
显存暴涨

全部为 0
可统一设定角色与边界,便于切换任务
文件大小
loss 异常
全部丢弃文件
每个任务有清晰的 system 或 tag 标识
中文专用
模型太大
逐字切片
提升多样性与鲁棒性,但需统一 prompt
复制多份
替换 attention
关闭 GPU
正确合并历史并应用角色标记与 loss mask
随机替换
更换模型
训练 1 轮
让模型分步推理后再给最终答案
替换 attention
Reduce + Accept
用高 temperature 即可
CoT 推理能力
鼓励幻觉
改 GPU
压缩权重
采样多条 CoT 推理并投票
降低显存
只给 Context
不同模型对比
Few-shot 引导效果下降
随便猜
加速推理
替换 tokenizer
在 prompt 中给出 JSON Schema 示例并要求严格遵循
重写历史
翻译成英文

关闭 CoT
用户输入覆盖或绕过系统指令
限制外部内容
压缩权重
禁用工具
随意增删关键约束
禁用 prompt
随机工具调用
语音清晰度
文本摘要/回答的召回度
更简单
Word
只看 ROUGE
固定随机种子、记录超参数与版本号
Python 版本
只跑 1 次
模型变好
准确性、流畅性、安全性、相关性
随意提交
倒序提交
只取最高分
评测集设计合理性与业务对齐度
随便输出
线上一定更差
暴露隐私
雷达图/分组柱状图
QPS
只提交报告
完全不固定
A/B 测试 + 关键业务指标
相信用户输入
回避问题
只看长尾
本地化模型权重 + 本地依赖 + 离线脚本
只看均值
人工全部放行
只看平均
明确截止时间并在报告中说明
完全冻结 W
INT8 缓存
梯度裁剪阈值
PeftModel.from_pretrained
验证 loss
Attention 的 q_proj
merge_adapter()
用于缓解过拟合
显存峰值尖刺
优化器状态减少
随机切分

Rerank 重排序
Recall@K
检索延迟
提升答案相关性
文本清洗
随机抽样
把文本映射为向量
System Prompt
检索召回
数据增强
TGI
加速 Attention 计算
模型参数量
提高 GPU 利用率
吞吐量
训练 Epoch 数
保留 top-k 候选
Q·K^T
增大 max_new_tokens
敏感信息
手机号
编辑距离
同义改写
过采样少数类
角色（user/assistant）
强制合并
Schema 一致
答非所问
标注一致性
示例尽可能长
让模型写出推理步骤
思考（Reasoning）
设定角色
Instruction
A/B 测试不同写法
<<<
多样优于重复
提升多步推理准确率
限定输出格式
lr_scheduler
按 max_length 截断
分层切分

空值率
字段重命名
Kappa 系数
混淆矩阵
字段重命名映射
按 max_length 截断
空回答
控制学习率
限制采样候选集
首字时延
调整推理参数
业务指标差异
KV cache 复用
多次重复
GPU 显存
使用更慢模型
准确性
JPEG
权重舍入误差
硬件兼容性
单文件部署
激活感知
跨平台
替换 Loss
精度 vs 速度
int4/int8 量化
量化支持
加载更大模型
nvidia-smi
requirements.txt
混合精度训练
低精度加载
torch 版本
把全部模型加载到 0 号卡
依赖解析能力
安装 NVIDIA 驱动
查看 nvidia-smi
显卡型号
固定随机种子
BLEU
模型 checkpoint 链接
有效性

基线模型
随机给分
折线图
核对字段
配对 t 检验
Embedding 层
Double Quantization
lora_dropout
设置 target_modules
使用 peft 库函数
显存占用
精度损失可控
AdaLoRA
LR Scheduler
推理时按需切换
递归字符切分
点积
支持多语言打分
向量维度
幻觉程度
检索召回
强调来源信息
IVF
保留段落边界
NDCG
降低每步计算量
DeepSpeed-Inference
减少 HBM 访问
统一推理
多 GPU 张量并行
top_k
使用 Flash Attention
请求排队过长
Sliding Window Attention
top_k
格式统一
邮箱地址
语义近似重复
句式变换
类别权重
自定义 JSON
纯表情回复

数据格式统一
态度友好度
全字段删除
示例顺序
Self-Consistency
Observation
角色身份设定
Output
模型重训
明确约束条件
示例与任务相关
多路径采样
限制输出格式
自定义 JSON
提高训练稳定性
比例合理
一致性
按需添加 few-shot
Kappa 系数
长度分布
格式统一
padding/truncation
重复样本去重
T 越大越发散
可与 temperature 组合
事实性评分
更换 RAG 召回
记录用户反馈
模型剪枝
多次重复
显存不足
相关性
降低 temperature
校准集偏差
跨 CPU/GPU 推理
对激活敏感度处理不同
推理速度
统一算子定义
FP16/INT8 量化
TensorRT Engine
时延
A/B 测试

误差补偿
释放中间变量
torch CUDA 是否可用
venv
混合精度训练
逐层加载
学习率
FSDP
环境复现
安装 cuDNN
使用混合精度
实验结果
公开数据集版本
F1
模型 checkpoint 链接
与业务目标对齐
A/B 实验
打分表设计
柱状图
核对文件命名
Bootstrap
学习率设置
Double Quantization
检索召回模块
相似度计算
Flash Attention
批处理大小
格式统一
数据格式
ReAct
示例的选择
长度分布
Instruction 拼接
吞吐量
top_k
便于部署
量化支持
cuDNN
Gradient Checkpointing
实验结果
超参配置记录
数据切分

lora_dropout
优化器状态
上传 Hub
AdaLoRA
可多任务切换 adapter
显存峰值
通过 PeftModel.from_pretrained
多路召回
索引大小
统一编码
用户问题
优化 Prompt 拼接
生成时延
表格
递归切分
模型并行度
复用前缀
端到端时延
Token 限流
减少 HBM 访问
多机多卡
多 LoRA 适配
位置编码外推
语种一致性
敏感词
文本归一化
场景标签
句式变换
Parquet
类别权重
掩码替换
Few-shot 示例
难度分布
多步规划
Observation
语言风格
输出校验
幻觉率
Output
Instruction 拼接
词云
时间序列切分

truncation 策略
样本多样性
重复样本
助手回答
Kappa 系数
显存峰值
top_k
更换 RAG 召回
关键指标
动态批处理
多次重复
相关性
降低 temperature
便于部署
GPTQ
CPU 推理
量化支持
统一算子
FP16/INT8 量化
校准集偏差
文档完善度
cuDNN
Gradient Checkpointing
venv
torch.cuda.is_available()
FSDP
逐层加载
锁定 Python 版本
启用混合精度
实验结果
公开数据集版本
实验配置
F1
A/B 实验
统一评分口径
热力图
核对文件命名

选项D	正确答案（必填）	归属题库（必填）
torch.device('gpu')	A	环境准备与模型资源管理
显存会自动释放,无需处理	B	环境准备与模型资源管理
提升模型输出精度	A	环境准备与模型资源管理
df.fillna(0)	A	数据探索与监督微调预处理
与 input_ids 相同	C	数据探索与监督微调预处理
对模型做蒸馏压缩	B	LoRA/QLoRA 微调训练
完全不需要反向传播	A	LoRA/QLoRA 微调训练
只比较生成速度	B	模型推理与效果对比
手动相加两个模型的参数	A	模型推理与效果对比
不释放显存以加快切换	B	模型推理与效果对比
替代 tokenizer	B	RAG 检索增强生成
加快分词速度	A	RAG 检索增强生成
训练集的大小	A	综合评估与结果提交
评测集无需参考答案	A	综合评估与结果提交
必须输出 JSON	A	提示工程实战
不使用任何约束	A	提示工程实战
直接丢弃文本	A	客服数据清洗与数据集融合
按文件大小去重	A	客服数据清洗与数据集融合
不做任何限制	A	客服数据清洗与数据集融合
三者吞吐完全相同	C	Transformers 并发推理
随机填充	A	Transformers 并发推理
增大模型参数量	A	模型量化与部署取舍

缩短推理文本长度	A	模型量化与部署取舍
只关注推理速度	A	模型量化与部署取舍
模型推理的平均速度	AB	环境准备与模型资源管理
模型权重应随提交压缩包打包	AC	环境准备与模型资源管理
直接调整模型权重	ABC	数据探索与监督微调预处理
统计缺失数量与原因后再决策	AD	数据探索与监督微调预处理
设计标签掩码使模型只对回答区域计算损失	ACD	数据探索与监督微调预处理
使用梯度累积	ABD	LoRA/QLoRA 微调训练
让每一步反向传播更省显存	BC	LoRA/QLoRA 微调训练
训练结束后及时释放显存	ABCD	LoRA/QLoRA 微调训练
两组使用不同的生成参数	AC	模型推理与效果对比
结论应与实际生成结果一致	BCD	模型推理与效果对比
相邻 chunk 保留一定重叠	BD	RAG 检索增强生成
文本向量化模型	CD	RAG 检索增强生成
返回顺序是否合理	ACD	RAG 检索增强生成
GPU 温度	AB	综合评估与结果提交
与实验数据一致的结论	ABCD	综合评估与结果提交
结构化输出约束	ABD	提示工程实战
示例输出与期望格式无需一致	AC	提示工程实战
提升复杂推理任务正确率	AD	提示工程实战
连续特殊符号	BD	客服数据清洗与数据集融合
是否包含 instruction 与 output 字段	CD	客服数据清洗与数据集融合
磁盘占用	AB	Transformers 并发推理
线程数越多吞吐越高且无上限	BC	Transformers 并发推理
无限制增大 batch_size	ABC	Transformers 并发推理

int4(NF4)量化	BCD	模型量化与部署取舍
	B	环境准备与模型资源管理
	B	数据探索与监督微调预处理
	B	LoRA/QLoRA 微调训练
	B	模型推理与效果对比
	B	RAG 检索增强生成
	B	综合评估与结果提交
	B	综合评估与结果提交
	B	提示工程实战
	A	客服数据清洗与数据集融合
	B	Transformers 并发推理
	B	模型量化与部署取舍
alpha	A	LoRA/QLoRA 微调训练
optimum	A	LoRA/QLoRA 微调训练
apply_lora	B	LoRA/QLoRA 微调训练
model.combine_weights()	C	LoRA/QLoRA 微调训练
lora_alpha	D	LoRA/QLoRA 微调训练
fp8	A	LoRA/QLoRA 微调训练
norm, bias	B	LoRA/QLoRA 微调训练
attention_dropout	C	LoRA/QLoRA 微调训练
model.save_pretrained()	D	LoRA/QLoRA 微调训练
将 batch size 调到 1	A	LoRA/QLoRA 微调训练
FP4 仅支持卷积	B	LoRA/QLoRA 微调训练
Mixture of Experts	C	LoRA/QLoRA 微调训练
单卡资源有限且需快速迭代多个业务适配	D	LoRA/QLoRA 微调训练
r 大小对效果无影响	A	LoRA/QLoRA 微调训练
GPU 利用率	B	LoRA/QLoRA 微调训练
Adapter 只在视觉模型中使用	C	LoRA/QLoRA 微调训练
进一步压缩量化常量本身的显存	D	LoRA/QLoRA 微调训练
重新启动训练	A	LoRA/QLoRA 微调训练
前者显存一定更少	B	LoRA/QLoRA 微调训练
重启整个 Python 进程	C	LoRA/QLoRA 微调训练
检查学习率与输入是否含异常字符	D	LoRA/QLoRA 微调训练
关闭评估	A	LoRA/QLoRA 微调训练
替代量化	B	LoRA/QLoRA 微调训练
LoRA 一定效果更好	C	LoRA/QLoRA 微调训练

LoRA 仅训练低秩矩阵与新增 bias,其余参数冻结且不存优化器状态	D	LoRA/QLoRA 微调训练
完全冻结 adapter	A	LoRA/QLoRA 微调训练
释放显存时	B	LoRA/QLoRA 微调训练
在 embedding 上做 LoRA	C	LoRA/QLoRA 微调训练
多 adapter 动态加载/卸载方案	D	LoRA/QLoRA 微调训练
强化对齐引导	A	RAG 检索增强生成
分词器	B	RAG 检索增强生成
生成训练数据	C	RAG 检索增强生成
避免语义被切断,提升召回质量	D	RAG 检索增强生成
Chroma	A	RAG 检索增强生成
训练轮数	B	RAG 检索增强生成
曼哈顿距离	C	RAG 检索增强生成
BGE	D	RAG 检索增强生成
作为 max_new_tokens 参数	A	RAG 检索增强生成
使用更小的模型	B	RAG 检索增强生成
检索速度加快	C	RAG 检索增强生成
对初筛结果二次精细排序以提升相关性	D	RAG 检索增强生成
随机采样	A	RAG 检索增强生成
越大越好无需控制	B	RAG 检索增强生成
提高 temperature	C	RAG 检索增强生成
按语义边界(段落/标题)切分并控制 token 数	D	RAG 检索增强生成
操作系统	A	RAG 检索增强生成
按页固定长度	B	RAG 检索增强生成
丢弃全部片段	C	RAG 检索增强生成
检索 + LLM 生成	D	RAG 检索增强生成
PQ 量化	A	RAG 检索增强生成
提升训练速度	B	RAG 检索增强生成
Python GC	C	RAG 检索增强生成
可表达更细粒度语义,但存储与检索开销增加	D	RAG 检索增强生成
Accuracy	A	RAG 检索增强生成
降低 top_k 到 1	B	RAG 检索增强生成
不需要 GPU	C	RAG 检索增强生成
检索 top_k 中片段排序不稳定或 temperature 较高	D	RAG 检索增强生成
无需索引	A	RAG 检索增强生成
提升训练速度	B	RAG 检索增强生成
model.decode(batch)	C	Transformers 并发推理
padding	D	Transformers 并发推理
替代 attention	A	Transformers 并发推理
标记关键词	B	Transformers 并发推理
流式保存 checkpoint	C	Transformers 并发推理
Python 解释器(CPython)	D	Transformers 并发推理
asyncio 串行	A	Transformers 并发推理
优化器选择	B	Transformers 并发推理
关闭 padding	C	Transformers 并发推理
CPU-GPU 之间的 host-to-device 传输	D	Transformers 并发推理
完全替代 attention 计算	A	Transformers 并发推理

embedding 维度变化	B	Transformers 并发推理
丢弃长 prompt	C	Transformers 并发推理
PagedAttention 高效管理 KV,提升吞吐	D	Transformers 并发推理
数据 schema	A	Transformers 并发推理
使用 FP16	B	Transformers 并发推理
切分到多进程	C	Transformers 并发推理
高吞吐生产级推理服务	D	Transformers 并发推理
改用 CPU	A	Transformers 并发推理
embedding 维度	B	Transformers 并发推理
1x1	C	Transformers 并发推理
缓解长序列/多请求场景下 KV cache 显存碎片	D	Transformers 并发推理
随机丢弃	A	Transformers 并发推理
改用 FP8	B	Transformers 并发推理
手动拆分权重	C	Transformers 并发推理
线性增长	D	Transformers 并发推理
关闭 mask	A	Transformers 并发推理
本地矩阵乘	B	Transformers 并发推理
提升 batch	C	Transformers 并发推理
torch.cuda.is_available()	D	环境准备与模型资源管
df -h	A	环境准备与模型资源管
torch.load_state	B	环境准备与模型资源管
加快数据加载	C	环境准备与模型资源管
减小 batch size 或序列长度	D	环境准备与模型资源管
flask	A	环境准备与模型资源管
导出 ONNX	B	环境准备与模型资源管
torch.show()	C	环境准备与模型资源管
权重加载报错或缺键/多余键	D	环境准备与模型资源管
换更小的 Python 版本	A	环境准备与模型资源管
切换 Linux	B	环境准备与模型资源管理
关闭屏幕	C	环境准备与模型资源管理
权重暂未加载,等首次前向时才真正加载	D	环境准备与模型资源管
FP16 + 关闭 KV cache	A	环境准备与模型资源管
cat transformers.txt	B	环境准备与模型资源管
清空磁盘	C	环境准备与模型资源管
部分张量未 .to(device) 与模型对齐	D	环境准备与模型资源管理
改用更小的模型结构	A	环境准备与模型资源管
日志	B	环境准备与模型资源管
改变 attention	C	环境准备与模型资源管
device_map='cpu' + 量化(load_in_4bit 关闭)或纯 CPU 推理	D	环境准备与模型资源管理
操作系统语言	A	环境准备与模型资源管
使用 TCP	B	环境准备与模型资源管
Python 版本错误	C	环境准备与模型资源管
BF16 加载或 4-bit 量化加载	D	环境准备与模型资源管
matplotlib	A	环境准备与模型资源管
强制 GPU 0	B	环境准备与模型资源管
磁盘空间不足	C	环境准备与模型资源管
NCCL 针对 GPU 高带宽优化,Gloo 偏 CPU/混合	D	环境准备与模型资源管

只加载一半层	A	环境准备与模型资源管
部署上线	B	客服数据清洗与数据集融合
加密压缩	C	客服数据清洗与数据集融合
JSON/CSV/Excel/数据库 dump	D	客服数据清洗与数据集融合
按时间倒序取前 N 条	A	客服数据清洗与数据集融合
转大写	B	客服数据清洗与数据集融合
删除一半数据	C	客服数据清洗与数据集融合
空对话、单字回复或纯寒暄无业务内容	D	客服数据清洗与数据集融合
广告文本	A	客服数据清洗与数据集融合
RGB	B	客服数据清洗与数据集融合
评估不准	C	客服数据清洗与数据集融合
使用 sentence embedding + 余弦相似度阈值	D	客服数据清洗与数据集融合
替换为 emoji	A	客服数据清洗与数据集融合
只取最后一句	B	客服数据清洗与数据集融合
压缩 zip	C	客服数据清洗与数据集融合
对少数类过采样或多数类降采样并评估	D	客服数据清洗与数据集融合
全部保留原样	A	客服数据清洗与数据集融合
训练步数减少	B	客服数据清洗与数据集融合
随机打散	C	客服数据清洗与数据集融合
易放大某些模式,建议去重或合并	D	客服数据清洗与数据集融合
随机位置	A	客服数据清洗与数据集融合
评估集挑简单的	B	客服数据清洗与数据集融合
随机替换	C	客服数据清洗与数据集融合
以工单事件为锚点拼接对话与处理结果	D	客服数据清洗与数据集融合
只支持英文	A	客服数据清洗与数据集融合
更换 tokenizer	B	客服数据清洗与数据集融合

训练完再处理	C	客服数据清洗与数据集融合
统一 schema 并按字段独立标注,缺失值置空	D	客服数据清洗与数据集融合
改用更大模型	A	客服数据清洗与数据集融合
改用图像分类模型	B	客服数据清洗与数据集融合
删除外部数据	C	客服数据清洗与数据集融合
降低显存占用,加速推理	D	模型量化与部署取舍
评估方法	A	模型量化与部署取舍
完全不可用	B	模型量化与部署取舍
bits=4	C	模型量化与部署取舍
AWQ 基于激活分布保护显著权重	D	模型量化与部署取舍
替换 attention	A	模型量化与部署取舍
embedding 检索	B	模型量化与部署取舍
评估	C	模型量化与部署取舍
8-bit 量化或仅 KV cache 量化	D	模型量化与部署取舍
标准 PyTorch	A	模型量化与部署取舍
只保留 adapter	B	模型量化与部署取舍
BitsAndBytesModel	C	模型量化与部署取舍
校准数据不足或不具代表性	D	模型量化与部署取舍
动态量化只支持 CPU	A	模型量化与部署取舍
数据不平衡	B	模型量化与部署取舍
不量化	C	模型量化与部署取舍
AWQ/GPTQ 而非简单 round-to-nearest	D	模型量化与部署取舍
加大 batch	A	模型量化与部署取舍
loss 更低	B	模型量化与部署取舍
切到 70B 模型	C	模型量化与部署取舍
精度、内存、延迟与功耗平衡	D	模型量化与部署取舍
关闭 attention	A	模型量化与部署取舍
替换模型	B	模型量化与部署取舍
无优化	C	模型量化与部署取舍
重新训练整个模型	D	模型量化与部署取舍
训练更快	A	模型量化与部署取舍
DeepSpeed 训练模式	B	模型量化与部署取舍
仅量化激活	C	模型量化与部署取舍
4-bit 量化 + 激活/KV 量化 + 设备 offload	D	模型量化与部署取舍
epochs	A	模型推理与效果对比
[0, 100]	B	模型推理与效果对比
禁用 attention	C	模型推理与效果对比
每步从概率最高的 k 个 token 中采样	D	模型推理与效果对比
改 learning rate	A	模型推理与效果对比
关闭所有采样	B	模型推理与效果对比
Python dict	C	模型推理与效果对比
生成时间显著增长且可能失控	D	模型推理与效果对比
关闭评估	A	模型推理与效果对比
固定 seed	B	模型推理与效果对比
关闭 attention	C	模型推理与效果对比
同一测试集、固定推理参数、多次重复取均值	D	模型推理与效果对比

完全随机	A	模型推理与效果对比
随机丢弃历史	B	模型推理与效果对比
batch size	C	模型推理与效果对比
低 temperature + 较小 top_p 或 greedy	D	模型推理与效果对比
固定 seed	A	模型推理与效果对比
只看困惑度	B	模型推理与效果对比
关闭 prompt	C	模型推理与效果对比
SamplingParams	D	模型推理与效果对比
seed 固定	A	模型推理与效果对比
关闭 GPU	B	模型推理与效果对比
只取第一个	C	模型推理与效果对比
离线集分布与线上不一致,或评测指标与业务指标脱节	D	模型推理与效果对比
不监控	A	模型推理与效果对比
随机	B	模型推理与效果对比
top_p=0	C	模型推理与效果对比
评测集、推理参数、对比指标、可复现脚本与结论	D	模型推理与效果对比
只看平均延迟	A	模型推理与效果对比
替换 attention	B	模型推理与效果对比
强化学习	C	数据探索与监督微调预处理
(instruction, response)	D	数据探索与监督微调预处理
数据增强	A	数据探索与监督微调预处理
SFT 训练时间更长	B	数据探索与监督微调预处理
替换 attention	C	数据探索与监督微调预处理
超过 max_length 时截断	D	数据探索与监督微调预处理
降低 batch	A	数据探索与监督微调预处理
改变 attention	B	数据探索与监督微调预处理
散点图(无轴)	C	数据探索与监督微调预处理
检查数据格式、标签对齐、学习率与 loss masking 是否正确	D	数据探索与监督微调预处理
直接喂入	A	数据探索与监督微调预处理
按日期	B	数据探索与监督微调预处理
训练无法启动	C	数据探索与监督微调预处理
人工 spot-check + 离线评测 + 线上 A/B	D	数据探索与监督微调预处理
tokenizer 失效	A	数据探索与监督微调预处理

随机	B	数据探索与监督微调预处理
替换 attention	C	数据探索与监督微调预处理
P50/P95/P99 长度及超长样本占比	D	数据探索与监督微调预处理
GPU 报错	A	数据探索与监督微调预处理
复制 1000 倍	B	数据探索与监督微调预处理
删除 tag	C	数据探索与监督微调预处理
支持中英文的多语言 tokenizer	D	数据探索与监督微调预处理
GPU 不够	A	数据探索与监督微调预处理
整段丢弃	B	数据探索与监督微调预处理
加快训练	C	数据探索与监督微调预处理
引入多样化 response 表达与改写	D	数据探索与监督微调预处理
提升速度	A	数据探索与监督微调预处理
继续满负荷训练	B	数据探索与监督微调预处理
随机 mask	C	数据探索与监督微调预处理
建立事实校验-人工审核-数据回灌闭环	D	数据探索与监督微调预处理
压缩权重	A	提示工程实战
使用大模型	B	提示工程实战
禁用 prompt	C	提示工程实战
设定模型角色、风格与边界	D	提示工程实战
Render + Action	A	提示工程实战
完全无需 prompt	B	提示工程实战
模型量化	C	提示工程实战
明确告知「只能基于给定知识回答,不知道就说不知道」	D	提示工程实战
关闭 attention	A	提示工程实战
降低显存	B	提示工程实战
关闭采样	C	提示工程实战
减少幻觉	D	提示工程实战
只给指令	A	提示工程实战
关闭监控	B	提示工程实战
loss 异常	C	提示工程实战
若不确定请回答「不知道」	D	提示工程实战
替换 attention	A	提示工程实战
提升 GPU	B	提示工程实战
禁止 JSON	C	提示工程实战
保留必要历史并明确 system 指令	D	提示工程实战
删除指令	A	提示工程实战

增加 batch	B	提示工程实战
网络中断	C	提示工程实战
完全相信用户输入	D	提示工程实战
GPU 加速	A	提示工程实战
随机调用	B	提示工程实战
明确输出格式	C	提示工程实战
将推理步骤表达为可执行代码以便外部计算	D	提示工程实战
禁用工具	A	提示工程实战
显存占用	B	综合评估与结果提交
GPU	C	综合评估与结果提交
使用语义向量衡量相似度	D	综合评估与结果提交
图片	A	综合评估与结果提交
只看 BERTScore	B	综合评估与结果提交
手动记录	C	综合评估与结果提交
结果摘要、对比分析、限制与展望	D	综合评估与结果提交
随机选指标	A	综合评估与结果提交
GPU 提升	B	综合评估与结果提交
只看大小写	C	综合评估与结果提交
在报告中明确指标并选择满足约束的方案	D	综合评估与结果提交
随便提交	A	综合评估与结果提交
只看均值	B	综合评估与结果提交
Python 版本	C	综合评估与结果提交
给出 JSON Schema 示例并明确校验失败回退	D	综合评估与结果提交
GPU 差异	A	综合评估与结果提交
减少页数	B	综合评估与结果提交
折线图(单一指标)	C	综合评估与结果提交
prompt 文件大小	D	综合评估与结果提交
只提交截图	A	综合评估与结果提交
随心情	B	综合评估与结果提交
不评估	C	综合评估与结果提交
在 system 中明确边界并加入分隔与校验	D	综合评估与结果提交
写无关内容	A	综合评估与结果提交
随机抽样	B	综合评估与结果提交
随机生成	C	综合评估与结果提交
通过有放回抽样估计指标分布并给出置信区间	D	综合评估与结果提交
随机替换敏感词	A	综合评估与结果提交
手动计时	B	综合评估与结果提交
每次随机	C	综合评估与结果提交
替换 W 为 B	ABC	LoRA/QLoRA 微调训练
全模型反量化	AB	LoRA/QLoRA 微调训练
Batch Size	AB	LoRA/QLoRA 微调训练
load_adapter	CD	LoRA/QLoRA 微调训练
训练步数上限	ACD	LoRA/QLoRA 微调训练
MLP 的 up_proj	CD	LoRA/QLoRA 微调训练
Trainer.load_model()	AB	LoRA/QLoRA 微调训练
用于替换 Attention 头	AC	LoRA/QLoRA 微调训练
optimizer 状态 OOM	CD	LoRA/QLoRA 微调训练
梯度计算量下降	CD	LoRA/QLoRA 微调训练
按词频切分	AB	RAG 检索增强生成

嵌入再训练	AC	RAG 检索增强生成
nDCG@K	CD	RAG 检索增强生成
召回质量	ACD	RAG 检索增强生成
减少 LLM 幻觉	CD	RAG 检索增强生成
格式标准化	CD	RAG 检索增强生成
Hash 索引	AB	RAG 检索增强生成
用于相似度检索	CD	RAG 检索增强生成
检索上下文	CD	RAG 检索增强生成
答案忠实度	CD	RAG 检索增强生成
模型蒸馏	ABD	Transformers 并发推理
TextCNN 推理	AC	Transformers 并发推理
降低显存占用	CD	Transformers 并发推理
GPU 算力	ACD	Transformers 并发推理
降低单请求时延	CD	Transformers 并发推理
首字时延	CD	Transformers 并发推理
批处理大小	ABD	Transformers 并发推理
通常提高生成质量	CD	Transformers 并发推理
与 V 加权求和	CD	Transformers 并发推理
使用更短序列	ABC	Transformers 并发推理
标准产品型号	ABC	客服数据清洗与数据集融
身份证号	CD	客服数据清洗与数据集融
语义向量相似度	CD	客服数据清洗与数据集融
回译	CD	客服数据清洗与数据集融
欠采样多数类	CD	客服数据清洗与数据集融
消息内容	CD	客服数据清洗与数据集融
重采样补全	ABD	客服数据清洗与数据集融
标签体系一致	CD	客服数据清洗与数据集融
空回答	CD	客服数据清洗与数据集融
样本多样性	CD	客服数据清洗与数据集融
示例随机选取	AB	提示工程实战
最终给出答案	CD	提示工程实战
行动（Action）	CD	提示工程实战
限制输出风格	CD	提示工程实战
Input	CD	提示工程实战
观察输出差异	CD	提示工程实战
<|im_end|>	ABCD	提示工程实战
由易到难效果更好	CD	提示工程实战
便于观察中间推理	CD	提示工程实战
对用户输入做转义	CD	提示工程实战
response	ABD	数据探索与监督微调预处
动态 padding 到 batch 内最大长度	CD	数据探索与监督微调预处
时间序列不能随机	BCD	数据探索与监督微调预处

重复率	CD	数据探索与监督微调预处
格式拼接	CD	数据探索与监督微调预处
双人标注一致性	CD	数据探索与监督微调预处
PS 抠图	ABC	数据探索与监督微调预处
嵌套字段展开	CD	数据探索与监督微调预处
保留头部+尾部	CD	数据探索与监督微调预处
超长 prompt	CD	数据探索与监督微调预处
控制批大小	AB	模型推理与效果对比
提高生成稳定性	CD	模型推理与效果对比
端到端时延	CD	模型推理与效果对比
改写 Prompt	CD	模型推理与效果对比
用户反馈差异	CD	模型推理与效果对比
量化	CD	模型推理与效果对比
使用相同测试集	BCD	模型推理与效果对比
网络 IO	CD	模型推理与效果对比
降低 temperature	ABD	模型推理与效果对比
流畅性	CD	模型推理与效果对比
MOV	AB	模型量化与部署取舍
激活值量化误差	CD	模型量化与部署取舍
生态成熟度	CD	模型量化与部署取舍
支持多种量化方案	CD	模型量化与部署取舍
按通道缩放	CD	模型量化与部署取舍
跨框架	CD	模型量化与部署取舍
增加 Epoch	AB	模型量化与部署取舍
显存 vs 吞吐	CD	模型量化与部署取舍
模型切分	BCD	模型量化与部署取舍
硬件适配	BCD	模型量化与部署取舍
调高 Batch Size	AB	环境准备与模型资源管理
nvcc --version	CD	环境准备与模型资源管理
pyproject.toml	CD	环境准备与模型资源管理
梯度累积	CD	环境准备与模型资源管理
device_map='auto'	CD	环境准备与模型资源管理
transformers 版本	CD	环境准备与模型资源管理
禁用梯度同步	AB	环境准备与模型资源管理
环境隔离能力	CD	环境准备与模型资源管理
安装 CUDA Toolkit	CD	环境准备与模型资源管理
降低 batch	CD	环境准备与模型资源管理
实验结果	ABCD	综合评估与结果提交
记录超参数	CD	综合评估与结果提交
ROUGE	CD	综合评估与结果提交
评估脚本	CD	综合评估与结果提交
与业务目标对齐	CD	综合评估与结果提交

消融实验	CD	综合评估与结果提交
统一口径培训	ABD	综合评估与结果提交
柱状图	CD	综合评估与结果提交
核对文件命名	CD	综合评估与结果提交
Bootstrap	CD	综合评估与结果提交
LayerNorm 层	AB	LoRA/QLoRA 微调训练
Paged Optimizers	ABCD	LoRA/QLoRA 微调训练
lora_target_modules	ABCD	LoRA/QLoRA 微调训练
选择学习率	BCD	LoRA/QLoRA 微调训练
必须重新训练	ABC	LoRA/QLoRA 微调训练
训练样本量	AB	LoRA/QLoRA 微调训练
完全不需要 GPU	ABC	LoRA/QLoRA 微调训练
全模型量化训练	ABC	LoRA/QLoRA 微调训练
Pretrain 重新训练基座	ABC	LoRA/QLoRA 微调训练
每次都合并到基座	ABC	LoRA/QLoRA 微调训练
随机字符级切分	ABC	RAG 检索增强生成
随机哈希匹配	ABC	RAG 检索增强生成
降低 Embedding 模型维度	AB	RAG 检索增强生成
检索表现	BCD	RAG 检索增强生成
训练时长	ABC	RAG 检索增强生成
重新训练基座	ABC	RAG 检索增强生成
使用最少上下文	ABC	RAG 检索增强生成
倒排索引（BM25）	ABC	RAG 检索增强生成
完全打散重排	ABC	RAG 检索增强生成
Loss 曲线	ABC	RAG 检索增强生成
替换 Attention 头	ABC	Transformers 并发推理
LoRA 训练	ABC	Transformers 并发推理
替换 Softmax	ABC	Transformers 并发推理
模型每批重新加载	ABC	Transformers 并发推理
反向传播优化	ABC	Transformers 并发推理
gradient_accumulation_steps	ABC	Transformers 并发推理
把所有 Attention 替换为全连接	ABC	Transformers 并发推理
训练数据不平衡	ABC	Transformers 并发推理
Dense 全连接替代	ABC	Transformers 并发推理
num_beams	ABCD	Transformers 并发推理
模型重新训练	ABC	客服数据清洗与数据集融
客户昵称	ABCD	客服数据清洗与数据集融
仅时间戳不同	ABC	客服数据清洗与数据集融
提升学习率	ABC	客服数据清洗与数据集融
调高 Batch Size	ABC	客服数据清洗与数据集融
图像文件	ABC	客服数据清洗与数据集融
正常业务问题	ABC	客服数据清洗与数据集融

替换底层模型	ABC	客服数据清洗与数据集融
显存占用	ABC	客服数据清洗与数据集融
随机打乱全部内容	AB	客服数据清洗与数据集融
示例要尽可能长	ABC	提示工程实战
无任何思维链	ABC	提示工程实战
Loss	ABC	提示工程实战
学习率	ABC	提示工程实战
Gradient	ABC	提示工程实战
更换硬件	AB	提示工程实战
任意自由发挥	ABC	提示工程实战
示例顺序无关	BC	提示工程实战
降低温度到 0	ABC	提示工程实战
删除所有用户输入	ABC	提示工程实战
BMP 图像	ABC	数据探索与监督微调预处
替换数据	ABC	数据探索与监督微调预处
只用训练集评估	ABC	数据探索与监督微调预处
显存峰值	ABC	数据探索与监督微调预处
随机修改标签	ABC	数据探索与监督微调预处
随机替换标签	ABC	数据探索与监督微调预处
模型准确率	ABC	数据探索与监督微调预处
删除所有样本	ABC	数据探索与监督微调预处
替换 Loss	ABC	数据探索与监督微调预处
把所有样本改写一遍	ABC	数据探索与监督微调预处
T 越小越发散	ABC	模型推理与效果对比
强制最大概率	ABC	模型推理与效果对比
训练 Loss	ABC	模型推理与效果对比
加大 Batch Size	ABC	模型推理与效果对比
只跑一次实验	ABC	模型推理与效果对比
重新训练基座	ABC	模型推理与效果对比
只跑一次	ABC	模型推理与效果对比
排队时延	CD	模型推理与效果对比
推理速度	ABC	模型推理与效果对比
提高 learning rate	ABC	模型推理与效果对比
训练 Loss 不一致	ABC	模型量化与部署取舍
需要联网授权	ABC	模型量化与部署取舍
都不支持 GPU	ABC	模型量化与部署取舍
生态成熟度	BCD	模型量化与部署取舍
完全取代训练流程	ABC	模型量化与部署取舍
完全替换 Loss	ABC	模型量化与部署取舍
PSD 文件	ABC	模型量化与部署取舍
硬件兼容性	BCD	模型量化与部署取舍
只看 Loss 曲线	ABC	模型量化与部署取舍

反向传播	ABC	模型量化与部署取舍
调高学习率	ABC	环境准备与模型资源管理
Photoshop 版本	ABC	环境准备与模型资源管理
记事本	ABC	环境准备与模型资源管理
梯度累积	CD	环境准备与模型资源管理
加载双份模型	ABC	环境准备与模型资源管理
CUDA 版本	ABD	环境准备与模型资源管理
单机单卡循环	ABC	环境准备与模型资源管理
环境激活/退出	BCD	环境准备与模型资源管理
安装 Photoshop	ABC	环境准备与模型资源管理
调高 epoch	ABC	环境准备与模型资源管理
显卡型号	ABCD	综合评估与结果提交
每次都换数据集	ABC	综合评估与结果提交
显存峰值	ABC	综合评估与结果提交
评估脚本	BCD	综合评估与结果提交
是否好看	ABC	综合评估与结果提交
无对照组	ABC	综合评估与结果提交
统一口径培训	ABC	综合评估与结果提交
热力图	BCD	综合评估与结果提交
重命名训练集	ABC	综合评估与结果提交
直接肉眼对比	ABC	综合评估与结果提交
target_modules 选择	ABCD	LoRA/QLoRA 微调训练
Paged Optimizers	ABCD	LoRA/QLoRA 微调训练
答案生成模块	ABCD	RAG 检索增强生成
Rerank 模型	ABCD	RAG 检索增强生成
模型量化	ABCD	Transformers 并发推理
解码策略	ABCD	Transformers 并发推理
质量评估	ABCD	客服数据清洗与数据集融
采样权重	ABCD	客服数据清洗与数据集融
System Prompt 设定	ABCD	提示工程实战
输出格式约束	ABCD	提示工程实战
类别分布	ABCD	数据探索与监督微调预处
数据切分	ABCD	数据探索与监督微调预处
显存占用	ABCD	模型推理与效果对比
repetition_penalty	ABCD	模型推理与效果对比
减少能耗	ABCD	模型量化与部署取舍
生态成熟度	ABCD	模型量化与部署取舍
PyTorch	ABCD	环境准备与模型资源管理
量化加载	ABCD	环境准备与模型资源管理
分析与结论	ABCD	综合评估与结果提交
环境依赖记录	ABCD	综合评估与结果提交
Instruction 模板拼接	ABCD	LoRA/QLoRA 微调训练

target_modules	ABCD	LoRA/QLoRA 微调训练
激活缓存	ABCD	LoRA/QLoRA 微调训练
必须丢弃 adapter	ABC	LoRA/QLoRA 微调训练
IA³	ABCD	LoRA/QLoRA 微调训练
效果一定更好	ABC	LoRA/QLoRA 微调训练
GPU 利用率	ABCD	LoRA/QLoRA 微调训练
必须重新训练基座	ABC	LoRA/QLoRA 微调训练
随机丢弃检索结果	ABC	RAG 检索增强生成
训练 Loss	ABC	RAG 检索增强生成
注入新 HTML 标签	ABC	RAG 检索增强生成
模型超参数	ABC	RAG 检索增强生成
随机删除上下文	ABC	RAG 检索增强生成
训练时长	ABC	RAG 检索增强生成
视频关键帧特征	ABCD	RAG 检索增强生成
整文档直接喂给 LLM	ABC	RAG 检索增强生成
训练 Epoch 数	ABC	Transformers 并发推理
完全禁用 Attention	ABC	Transformers 并发推理
训练 Loss	ABC	Transformers 并发推理
完全不限流	ABC	Transformers 并发推理
提升模型参数量	ABC	Transformers 并发推理
离线离线再离线	ABC	Transformers 并发推理
反向传播更新	ABC	Transformers 并发推理
训练集变大	ABC	Transformers 并发推理
显存占用	ABC	客服数据清洗与数据集融
高质量专业回答	ABC	客服数据清洗与数据集融
替换底层模型	ABC	客服数据清洗与数据集融
学习率	ABC	客服数据清洗与数据集融
提高 Loss	ABC	客服数据清洗与数据集融
BMP	ABC	客服数据清洗与数据集融
随机丢弃所有样本	ABC	客服数据清洗与数据集融
原始数据公开	ABC	客服数据清洗与数据集融
强制模型输出 Loss	ABC	提示工程实战
示例越长越好	ABC	提示工程实战
简单事实查询	ABC	提示工程实战
Loss	ABC	提示工程实战
训练轮次	ABC	提示工程实战
让模型自己写 Loss	ABC	提示工程实战
显存占用	ABC	提示工程实战
GPU 编号	ABC	提示工程实战
替换底层硬件	ABC	数据探索与监督微调预处
UI 设计稿	ABC	数据探索与监督微调预处
全部用作训练	ABC	数据探索与监督微调预处

训练 Loss	ABC	数据探索与监督微调预处
显卡型号	ABC	数据探索与监督微调预处
格式规范的标准样本	ABC	数据探索与监督微调预处
模型版本号	ABC	数据探索与监督微调预处
随机修改标签	ABC	数据探索与监督微调预处
训练时长	ABC	模型推理与效果对比
num_beams	ABC	模型推理与效果对比
加大 batch	ABC	模型推理与效果对比
完全离线评测	ABC	模型推理与效果对比
反向传播更新	ABC	模型推理与效果对比
只跑一次	ABC	模型推理与效果对比
显存峰值	ABC	模型推理与效果对比
提高 learning rate	ABC	模型推理与效果对比
提升 Loss	ABC	模型量化与部署取舍
AWQ	ABCD	模型量化与部署取舍
需要联网授权	ABC	模型量化与部署取舍
学习率	ABC	模型量化与部署取舍
自动反向传播	ABC	模型量化与部署取舍
删除 Loss	ABC	模型量化与部署取舍
数据标注噪声	ABC	模型量化与部署取舍
训练集大小	ABC	模型量化与部署取舍
PyTorch / Transformers	ABCD	环境准备与模型资源管理
量化加载	ABCD	环境准备与模型资源管理
Photoshop	ABC	环境准备与模型资源管理
ipconfig	ABC	环境准备与模型资源管理
Tensor Parallel	ABCD	环境准备与模型资源管理
完全在 CPU 推理	ABC	环境准备与模型资源管理
随机改版本	ABC	环境准备与模型资源管理
调高 epoch	ABC	环境准备与模型资源管理
分析与结论	ABCD	综合评估与结果提交
随意改环境	ABC	综合评估与结果提交
显卡 BIOS	ABC	综合评估与结果提交
显存峰值	ABC	综合评估与结果提交
无对照组	ABC	综合评估与结果提交
随机打 0 分	ABC	综合评估与结果提交
PPT 海报	ABC	综合评估与结果提交
重命名训练集	ABC	综合评估与结果提交
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练

	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	LoRA/QLoRA 微调训练
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	RAG 检索增强生成
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	Transformers 并发推理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	环境准备与模型资源管理
	A	客服数据清洗与数据集融合
	A	客服数据清洗与数据集融合
	A	客服数据清洗与数据集融合
	A	客服数据清洗与数据集融合
	A	客服数据清洗与数据集融合
	A	客服数据清洗与数据集融合

	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	多服数据清洗与数据集融合
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	A	模型量化与部署取舍
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	LoRA/QLoRA 微调训练
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	RAG 检索增强生成
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	Transformers 并发推理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	环境准备与模型资源管理
	B	多服数据清洗与数据集融合
	B	多服数据清洗与数据集融合
	B	多服数据清洗与数据集融合
	B	多服数据清洗与数据集融合

	B	克服数据清洗与数据集融合
	B	克服数据清洗与数据集融合
	B	克服数据清洗与数据集融合
	B	克服数据清洗与数据集融合
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型量化与部署取舍
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	模型推理与效果对比
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	数据探索与监督微调预处理
	B	提示工程实战
	B	提示工程实战
	B	提示工程实战
	B	提示工程实战
	B	提示工程实战
	B	提示工程实战
	B	提示工程实战
	B	综合评估与结果提交
	B	综合评估与结果提交
	B	综合评估与结果提交
	B	综合评估与结果提交
	B	综合评估与结果提交

归属知识点（必填）	难度（必填）
CUDA 与 GPU 环境	极简单
显存管理	普通
模型加载	难
缺失值处理	极简单
标签掩码	普通
LoRA 原理	极简单
QLoRA	普通
对比实验	极简单
LoRA 推理	普通
显存管理	普通
RAG 原理	极简单
向量检索	简单
BLEU	简单
评测集划分	普通
Zero-shot	极简单
结构化输出	普通
对话解析	简单
去重	普通
长度过滤	普通
推理模式	简单
批处理	难
量化收益	简单

双重量化	普通
部署取舍	普通
环境自检	普通
模型资源管理	普通
数据探索	普通
缺失值处理	难
标签掩码	极难
低显存训练	普通
梯度累积	难
训练流程	普通
测试集设计	简单
对比分析	难
文本切分	普通
RAG 组件	难
检索验证	难
评估指标	极难
评估报告	普通
提示技巧	普通
Few-shot	简单
CoT	难
去噪	难
JSONL 校验	简单
性能指标	普通
线程并发	简单
OOM 处理	普通

量化方案	普通
异常处理	简单
标签掩码	难
LoRA 原理	简单
对比实验	普通
RAG 原理	极难
评估指标	简单
评估报告	普通
提示技巧	普通
去噪	普通
线程并发	极难
量化权衡	难
LoRA 原理	极简单
QLoRA 量化	极简单
PEFT 库	极简单
LoRA 权重合并	简单
LoRA 超参数	简单
QLoRA 量化	简单
target_modules	简单
LoRA 超参数	简单
LoRA 权重合并	简单
训练超参数	普通
QLoRA 量化	普通
显存优化	普通
LoRA 选型	普通
LoRA 超参数	普通
效果评估	普通
LoRA 原理	普通
QLoRA 量化	普通
LoRA 权重合并	普通
target_modules	普通
PEFT 库	普通
异常处理	普通
训练超参数	难
QLoRA 量化	难
LoRA 原理	难

LoRA 原理	难
多任务适配	难
显存优化	难
LoRA 原理	极难
多任务适配	极难
RAG 原理	极简单
Embedding	极简单
向量数据库	极简单
文档切片	简单
向量数据库	简单
检索参数	简单
检索参数	简单
Embedding	简单
Prompt 融合	简单
效果调优	普通
文档切片	普通
Rerank	普通
检索方式	普通
Rerank	普通
效果调优	普通
文档切片	普通
效果调优	普通
文档切片	普通
Prompt 融合	普通
系统延迟	普通
向量数据库	普通
Embedding	普通
向量数据库	难
Embedding	难
效果评估	难
效果调优	难
检索方式	难
效果调优	难
RAG 原理	极难
检索方式	极难
batch 推理	极简单
padding	极简单
KV cache	简单
attention mask	简单
流式推理	简单
并发模型	简单
并发模型	简单
padding	简单
batch 推理	普通
并发模型	普通
attention 计算	普通

KV cache	普通
batch 推理	普通
并发模型	普通
推理参数	普通
显存管理	普通
并发模型	普通
并发模型	普通
并发模型	普通
推理参数	普通
attention mask	普通
KV cache	难
并发模型	难
推理优化	难
并发模型	难
KV cache	难
推理优化	难
并发模型	极难
KV cache	极难
CUDA 与 GPU 环境	极简单
显存管理	极简单
模型加载	极简单
模型加载	简单
显存管理	简单
环境依赖	简单
模型加载	简单
CUDA 与 GPU 环境	简单
异常处理	简单
模型加载	普通
异常处理	普通
异常处理	普通
模型加载	普通
显存管理	普通
环境依赖	普通
显存管理	普通
异常处理	普通
显存管理	普通
模型加载	普通
模型加载	普通
CUDA 与 GPU 环境	普通
环境自检	普通
模型加载	难
异常处理	难
显存管理	难
环境依赖	难
模型加载	难
CUDA 与 GPU 环境	难
环境依赖	极难

模型加载	极难
数据清洗	极简单
数据清洗	极简单
数据格式	极简单
数据清洗	简单
数据清洗	简单
数据集融合	简单
数据清洗	简单
数据清洗	简单
数据格式	简单
数据集融合	普通
数据清洗	普通
数据清洗	普通
数据格式	普通
数据清洗	普通
数据平衡	普通
数据清洗	普通
数据质量	普通
数据集融合	普通
数据清洗	普通
数据格式	普通
数据质量	普通
数据清洗	普通
数据集融合	难
数据清洗	难
数据质量	难

数据清洗	难
数据格式	难
数据清洗	难
数据格式	极难
数据集融合	极难
量化原理	极简单
训练后量化	极简单
量化精度损失	简单
4-bit 量化	简单
AWQ/GPTQ	简单
模型导出	简单
推理引擎	简单
部署格式	简单
量化精度损失	普通
部署格式	普通
量化部署	普通
GPTQ	普通
量化精度损失	普通
量化原理	普通
量化原理	普通
推理引擎	普通
量化精度损失	普通
异常处理	普通
模型导出	普通
部署格式	普通
部署格式	普通
量化部署	难
KV cache 量化	难
推理引擎	难
部署格式	难
量化精度损失	难
推理引擎	难
AWQ/GPTQ	极难
部署格式	极难
推理参数	极简单
推理参数	极简单
推理参数	极简单
推理参数	简单
推理参数	简单
推理参数	简单
推理调用	简单
推理参数	简单
效果评估	简单
推理参数	普通
推理参数	普通
效果评估	普通

推理参数	普通
推理调用	普通
效果评估	普通
推理参数	普通
推理参数	普通
效果评估	普通
效果评估	普通
推理参数	普通
推理参数	普通
推理参数	普通
推理参数	难
效果评估	难
效果评估	难
推理参数	难
推理参数	难
效果评估	难
效果评估	极难
推理参数	极难
SFT 原理	极简单
SFT 数据格式	极简单
tokenizer	极简单
SFT 原理	简单
padding/truncation	简单
padding/truncation	简单
loss masking	简单
instruction template	简单
EDA	简单
SFT 训练排查	普通
padding/truncation	普通
数据清洗	普通
SFT 数据质量	普通
数据质量评估	普通
数据质量评估	普通

padding/truncation	普通
instruction template	普通
EDA	普通
SFT 数据质量	普通
数据清洗	普通
instruction template	普通
tokenizer	普通
loss masking	难
padding/truncation	难
SFT 数据格式	难
SFT 数据质量	难
数据质量评估	难
SFT 训练排查	难
instruction template	极难
数据质量评估	极难
Few-shot	极简单
Zero-shot	极简单
CoT	简单
System Prompt	简单
ReAct	简单
角色设定	简单
CoT	简单
约束 Prompt	简单
Few-shot	普通
Few-shot	普通
CoT	普通
约束 Prompt	普通
Prompt 结构	普通
Prompt 优化	普通
Few-shot	普通
约束 Prompt	普通
CoT	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
Prompt 优化	普通

CoT	难
安全 Prompt	难
安全 Prompt	难
ReAct	难
Prompt 结构	难
Prompt 优化	难
CoT	极难
ReAct	极难
BLEU	极简单
ROUGE	极简单
BERTScore	简单
提交格式	简单
评估指标	简单
可复现性	简单
报告撰写	简单
评估指标	简单
评估指标	普通
评估指标	普通
提交格式	普通
可复现性	普通
评估指标	普通
报告撰写	普通
报告撰写	普通
评估指标	普通
报告撰写	普通
报告撰写	普通
评估指标	普通
可复现性	普通
可复现性	普通
评估指标	普通
可复现性	难
报告撰写	难
评估指标	难
提交格式	难
评估指标	难
评估指标	难
评估指标	极难
报告撰写	极难
LoRA 原理	简单
QLoRA 量化	简单
LoRA 超参数	难
PEFT 库	普通
训练规范	简单
target_modules	普通
PEFT 库	普通
LoRA 超参数	难
QLoRA 量化	难
LoRA 原理	简单
文档切片	简单

Rerank 重排序	简单
RAG 评估	普通
向量检索	普通
Rerank 重排序	简单
RAG 架构	普通
向量检索	简单
Embedding 模型	极简单
Prompt 拼接	简单
RAG 评估	普通
KV cache	简单
并发模型	简单
Flash Attention	普通
性能指标	简单
动态批处理	普通
推理引擎	简单
推理参数	普通
推理参数	普通
Attention 优化	极简单
OOM处理	普通
数据清洗	简单
隐私脱敏	极简单
数据去重	普通
数据增强	普通
样本均衡	普通
数据格式	简单
异常处理	普通
数据格式	普通
数据清洗	简单
数据清洗	普通
Few-shot	简单
CoT	简单
ReAct	普通
System Prompt	简单
instruction	简单
提示词调优	普通
Prompt 结构	普通
Few-shot	难
CoT	简单
提示词调优	难
数据格式	简单
padding/truncation	普通
数据切分	普通

数据质量评估	简单
instruction	普通
样本标注	普通
数据质量评估	简单
数据集融合	普通
padding/truncation	普通
数据质量评估	简单
temperature/top_p	简单
temperature/top_p	简单
效果评估	简单
效果调优	简单
A/B 测试	普通
推理优化	普通
对比实验	普通
推理优化	普通
对比实验	简单
效果评估	简单
GGUF	简单
量化精度损失	普通
部署格式	普通
GGUF	普通
GPTQ/AWQ	难
ONNX	普通
TensorRT	普通
部署格式	简单
部署策略	普通
推理引擎	简单
显存管理	简单
CUDA 与 GPU 环境	简单
依赖管理	简单
显存优化	普通
模型加载	普通
环境依赖	简单
显存管理	普通
包管理	普通
CUDA 与 GPU 环境	简单
显存管理	简单
报告撰写	简单
可复现性	简单
BLEU/ROUGE	简单
结果提交	简单
评估指标	普通

实验设计	普通
人工评估	普通
报告撰写	极简单
结果提交	简单
评估指标	难
target_modules	简单
QLoRA 量化	简单
LoRA 超参数	简单
LoRA 训练流程	简单
LoRA 权重合并	简单
LoRA 超参数	简单
QLoRA 量化	简单
PEFT 库	极简单
LoRA 训练流程	普通
LoRA 原理	难
文档切片	普通
向量检索	简单
Rerank 重排序	普通
Embedding 模型	简单
RAG 评估	普通
RAG 架构	简单
Prompt 拼接	难
向量检索	难
文档切片	普通
RAG 评估	难
KV cache	简单
并发模型	简单
Flash Attention	普通
动态批处理	普通
推理引擎	难
推理参数	简单
Attention 优化	普通
并发模型	普通
KV cache	难
推理参数	普通
数据清洗	简单
隐私脱敏	极简单
数据去重	普通
数据增强	普通
样本均衡	普通
数据格式	简单
异常处理	普通

数据格式	普通
数据清洗	普通
异常处理	难
Few-shot	简单
CoT	简单
ReAct	普通
System Prompt	简单
instruction	简单
提示词调优	普通
Prompt 结构	简单
Few-shot	普通
CoT	普通
System Prompt	难
数据格式	简单
padding/truncation	简单
数据切分	普通
数据质量评估	简单
instruction	普通
样本标注	普通
数据质量评估	简单
数据集融合	普通
padding/truncation	简单
数据质量评估	普通
temperature/top_p	简单
temperature/top_p	普通
效果评估	简单
效果调优	普通
A/B 测试	普通
推理优化	普通
对比实验	普通
推理优化	普通
效果评估	简单
效果调优	难
量化精度损失	普通
GGUF	普通
GPTQ/AWQ	难
部署格式	简单
ONNX	普通
TensorRT	普通
部署格式	简单
推理引擎	普通
量化精度损失	普通

GPTQ/AWQ	难
显存管理	简单
CUDA 与 GPU 环境	简单
依赖管理	极简单
显存优化	普通
模型加载	普通
环境依赖	简单
显存管理	普通
包管理	普通
CUDA 与 GPU 环境	简单
显存管理	简单
报告撰写	简单
可复现性	简单
BLEU/ROUGE	简单
结果提交	简单
评估指标	普通
实验设计	普通
人工评估	普通
报告撰写	极简单
结果提交	简单
评估指标	难
LoRA 原理	简单
QLoRA 量化	简单
RAG 架构	简单
向量检索	普通
推理优化	普通
推理参数	普通
数据清洗	简单
数据格式	普通
Prompt 结构	简单
instruction	普通
数据质量评估	简单
数据格式	普通
效果评估	普通
temperature/top_p	简单
量化精度损失	简单
部署格式	普通
CUDA 与 GPU 环境	简单
显存优化	普通
报告撰写	简单
可复现性	普通
LoRA 训练流程	简单

LoRA 超参数	简单
QLoRA 量化	普通
LoRA 权重合并	普通
PEFT 库	普通
LoRA 原理	普通
LoRA 训练流程	简单
LoRA 权重合并	普通
向量检索	普通
向量检索	普通
RAG 架构	简单
Prompt 拼接	简单
RAG 评估	普通
RAG 评估	普通
Embedding 模型	难
文档切片	简单
并发模型	普通
KV cache	普通
并发模型	简单
推理引擎	普通
Flash Attention	普通
推理引擎	简单
推理引擎	普通
推理参数	难
数据清洗	简单
异常处理	简单
数据格式	普通
样本标注	普通
数据增强	普通
数据格式	简单
样本均衡	普通
隐私脱敏	普通
Prompt 结构	简单
Few-shot	普通
CoT	普通
ReAct	普通
System Prompt	简单
提示词调优	难
提示词调优	普通
instruction	简单
数据格式	简单
数据质量评估	简单
数据切分	普通

padding/truncation	普通
数据质量评估	普通
数据质量评估	简单
instruction	简单
样本标注	普通
效果评估	简单
temperature/top_p	简单
效果调优	普通
A/B 测试	普通
推理优化	普通
对比实验	普通
效果评估	简单
效果调优	难
量化精度损失	简单
GPTQ/AWQ	普通
GGUF	普通
部署格式	简单
ONNX	普通
TensorRT	普通
量化精度损失	普通
部署格式	普通
CUDA 与 GPU 环境	简单
显存优化	普通
依赖管理	极简单
CUDA 与 GPU 环境	简单
显存管理	普通
模型加载	普通
包管理	简单
显存管理	简单
报告撰写	简单
可复现性	普通
结果提交	普通
BLEU/ROUGE	简单
实验设计	普通
人工评估	普通
报告撰写	简单
结果提交	简单
LoRA 原理	极简单
QLoRA 量化	极简单
LoRA 超参数	简单
LoRA 权重合并	简单
LoRA 超参数	普通
QLoRA 量化	普通

显存优化	普通
异常处理	普通
LoRA 原理	难
多任务适配	难
显存优化	极难
RAG 原理	极简单
向量数据库	极简单
检索参数	简单
检索方式	简单
Rerank	普通
文档切片	普通
Embedding	普通
效果调优	普通
系统延迟	普通
文档切片	普通
RAG 原理	难
效果评估	难
检索方式	极难
batch 推理	极简单
KV cache	简单
attention mask	简单
并发模型	简单
attention 计算	普通
KV cache	普通
并发模型	普通
padding	普通
并发模型	普通
推理优化	普通
KV cache	难
KV cache	难
KV cache	难
并发模型	极难
CUDA 与 GPU 环境	极简单
显存管理	极简单
模型加载	简单
显存管理	简单
CUDA 与 GPU 环境	简单
模型加载	普通
模型加载	普通
异常处理	普通
模型加载	普通
环境依赖	难
模型加载	难
环境自检	难
环境依赖	极难
数据清洗	极简单
数据格式	极简单
数据清洗	简单
数据集融合	简单
数据清洗	普通
数据格式	普通

数据质量	普通
数据清洗	普通
数据质量	普通
数据清洗	普通
数据清洗	难
数据格式	难
数据集融合	极难
训练后量化	极简单
量化原理	简单
4-bit 量化	简单
模型导出	简单
部署格式	普通
推理引擎	普通
量化原理	普通
量化原理	普通
AWQ/GPTQ	普通
量化精度损失	普通
LoRA 原理	极简单
QLoRA 量化	简单
LoRA 超参数	简单
LoRA 权重合并	简单
LoRA 原理	普通
QLoRA 量化	普通
LoRA 权重合并	普通
显存优化	极难
RAG 原理	极简单
向量数据库	简单
检索参数	简单
检索方式	普通
文档切片	普通
系统延迟	普通
RAG 原理	难
检索方式	极难
batch 推理	极简单
KV cache	普通
attention 计算	简单
并发模型	普通
padding	普通
推理优化	难
并发模型	极难
CUDA 与 GPU 环境	简单
显存管理	简单
模型加载	简单
显存管理	普通
模型加载	普通
模型加载	普通
环境依赖	极难
数据清洗	极简单
数据清洗	简单
数据集融合	简单
数据清洗	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通

数据集融合	普通
数据格式	普通
数据质量	普通
数据集融合	极难
训练后量化	简单
4-bit 量化	简单
4-bit 量化	简单
模型导出	普通
部署格式	普通
量化原理	普通
量化部署	普通
推理参数	简单
推理参数	简单
推理参数	普通
推理参数	简单
效果评估	普通
推理参数	普通
效果评估	难
SFT 原理	极简单
tokenizer	简单
padding/truncation	简单
padding/truncation	普通
loss masking	普通
SFT 训练排查	普通
SFT 数据质量	难
Few-shot	简单
CoT	简单
System Prompt	普通
ReAct	普通
Prompt 结构	普通
Prompt 结构	普通
安全 Prompt	难
BLEU	极简单
ROUGE	简单
BERTScore	简单
评估指标	普通
提交格式	普通