Artificial Intelligence PhD Qualifying Examination

模拟考试 10：参考解答与评分要点

每个 Section 为 33 分；考生选四个 Section 中的三个作答，姓名与学号另计 1 分。以下给出完整参考推导。系统设计题中的方案均为 representative full-credit design；其他满足题面核心约束、论证完整且技术可行的方案同等给分。

A. 机器学习理论：参考解答 [33 points]

Part I. $\boxed{\text{(i)a, (ii)a, (iii)a, (iv)a, (v)a, (vi)a}}$。

Part II.

(a) (a) 若 clean prediction correct，noise 后出错概率为 η；若 clean prediction wrong，noise 后出错概率为 $1-\eta$。所以

$$\widetilde{L}(h) = \eta[1 - L(h)] + (1-\eta)L(h) = \eta + (1-2\eta)L(h).$$

(b) 因 $1-2\eta > 0$，正 affine transformation 保持 risk ordering，且不要求 infimum 必须由某个 hypothesis 取到。直接有

$$\inf_{g\in\mathcal{H}} \widetilde{L}(g) = \eta + (1-2\eta) \inf_{g\in\mathcal{H}} L(g),$$

所以

$$\widetilde{L}(h) - \inf_g \widetilde{L}(g) = (1-2\eta)\left[L(h) - \inf_g L(g)\right].$$

故 clean excess 至多 $\boxed{\gamma/(1-2\eta)}$。当 $\eta \uparrow 1/2$，denominator 趋零，同一 noisy-risk precision 对 clean risk 的控制退化。

Rubric: 两种 clean cases 2 分；identity 2 分；ordering/infimum 1 分；excess conversion 2 分；limit 1 分。

(b) 固定 h。若 $L(h) > \varepsilon_h$，它在全部 sample 上 consistent 的概率

$$(1 - L(h))^m \le e^{-mL(h)} < e^{-m\varepsilon_h}.$$

取

$$\varepsilon_h = \frac{\log(1/\pi(h)) + \log(1/\delta)}{m}$$

后，上式等于 $\pi(h)\delta$。对 countable $\mathcal{H}$ union bound，存在违反相应 bound 且 consistent 的 h 的概率至多 $\sum_h \pi(h)\delta = \delta$。所以结论 simultaneously 成立。

Rubric: fixed-h survival 2 分；hypothesis-dependent threshold 2 分；weighted union 2 分；simultaneous conclusion 1 分。

(c) 下界：取 $2d$ 点 $\{\pm e_j : j \in [d]\}$。对任意要标为 1 的 subset，可逐 coordinate 选择 lower/upper endpoints：是否包含 $-e_j$ 决定 lower endpoint 是否越过 -1，是否包含 e_j 决定 upper endpoint 是否越过 1；所有其他 coordinates 为 0，可使各选择独立。Empty positive set 用远离这些点的小 box。因此这 $2d$ 点被 shattered。

上界：任取 $2d+1$ 个 distinct points。对每个 coordinate 选一个 minimum point 与一个 maximum point，共至多 $2d$ 个 selected extrema，所以有一点 x° 未选。任何包含全部 selected extrema 的 axis-aligned box 必须包含整个 sample 的 coordinatewise bounding box，从而也包含 x°。因此“extrema 标 1、x° 标 0”的 labeling 无法实现。故

$$\boxed{\mathrm{VCdim}(\mathcal{B}_d) = 2d.}$$

Rubric: $2d$ 点 1 分；independent endpoint construction 3 分；extrema selection 2 分；upper-bound labeling/结论 2 分。

(d) Strong convexity 等价于 $G(w) = F(w) - (\lambda/2)\,\|w\|^2$ convex。Jensen 用于 G：

$$\mathbb{E}F(W) - \frac{\lambda}{2}\mathbb{E}\,\|W\|^2 \geq F(\mathbb{E}W) - \frac{\lambda}{2}\,\|\mathbb{E}W\|^2 .$$

移项并使用 variance identity $\mathbb{E}\,\|W\|^2 - \|\mathbb{E}W\|^2 = \mathbb{E}\,\|W - \mathbb{E}W\|^2$，即得题式。故 parameter randomization 的 expected objective 至少比 mean parameter 的 objective 多一个由 parameter variance 与 curvature 控制的 gap；这不声称非凸 neural objectives 也满足同一式。

Rubric: convex transform 2 分；Jensen 2 分；variance identity 1 分；解释/caveat 2 分。

B. 深度学习与强化学习：参考解答 [33 points]

(a) **Decoder-only Transformer**。 [9 points]

(a) $Q, K, V \in \mathbb{R}^{T\times d_k}$，score/attention matrix 为 $T \times T$。可令 $M_{ij} = 0$ 当 $j \le i$，$M_{ij} = -\infty$ 当 $j > i$，并计算 $A = \mathrm{softmax}_{\mathrm{row}}(QK^\top/\sqrt{d_k} + M)$。softmax 对每个 query 在 key 轴归一化。

(b) 训练时真实 prefix 全部已知，masked matrix operations 可同时计算所有位置的 logits 与 next-token losses。采样时 x_t 必须先生成，才能成为未来条件，故时间步严格顺序。矩阵内部并行不等于跨生成步并行。

(c) 每层训练主项 $O(Td^2 + T^2d)$。若第 t 步重算整个 prefix，总代价 $O(T^2d^2 + T^3d)$。KV cache 保存旧 token 的 keys/values，第 t 步约 $O(d^2 + td)$，总计 $O(Td^2 + T^2d)$；每层 cache 为 $O(Td)$（多头常数已吸收）。

Rubric: mask/dimensions/axis 3 分；teacher forcing 与 sequential sampling 3 分；三种 complexity/cache memory 3 分。

(b) **Autoregressive likelihood**。 [14 points]

(a) 从最后一维开始求和：

$$\sum_{x_{1:T}} \prod_{t=1}^{T} p(x_t \mid x_{<t}) = \sum_{x_{1:T-1}} \prod_{t=1}^{T-1} p(x_t \mid x_{<t}) \underbrace{\sum_{x_T} p(x_T \mid x_{<T})}_{1}$$

$$= \cdots = 1.$$

链式分解是概率恒等式；近似来自有限模型族、上下文、数据与优化，而非 factorization 本身。

(b) 令 $m_{nt} \in \{0, 1\}$ 指示非 padding target，$N_{\mathrm{tok}} = \sum_{n,t} m_{nt}$，则 token-average NLL 可写

$$\mathcal{L} = -\frac{1}{N_{\mathrm{tok}}} \sum_{n,t} m_{nt} \log p_\theta(x_{nt} \mid x_{n,<t}).$$

也可先逐序列平均再对序列平均，但两者对不同长度序列权重不同，必须明确且梯度中只除一次。

(c) Exposure bias 指训练 prefix 来自数据、推理 prefix 来自模型，后者的早期错误会改变未来条件分布。Teacher forcing 的 loss 仍精确等于 joint NLL 的条件项之和；问题来自有限样本/模型错设和 rollout 分布变化，不是链式公式错误。

(d) 从 BOS 与 prompt 开始，逐步计算 $p_\theta(\cdot \mid x_{<t})$，按指定 rule 采样/选择 x_t，直到 EOS 或长度上限。高 likelihood 不保证最好感知质量，因为 forward-KL/NLL 对所有数据模式加权而感知指标不同；解码策略会改变样本；有限测试 NLL 也不测 factuality/diversity。任两项。

(e) 除以长度得到平均 token log-score，可减少长序列因累加更多负 log-prob 而受罚，常用于 ranking/decoding。但 $p(x)^{1/T}$ 在不同长度序列上通常不自动归一化，因此不是原 joint probability，也不能在未给 normalizer 时称为新概率模型。

Rubric: iterated normalization/非近似 4 分；masked loss 与 convention 3 分；exposure 2 分；sampling/两个 distinction 3 分；length normalization 2 分。

(c) **Sequence policy gradient**。 [10 points]

(a) 状态 $S_t = (x, y_{<t})$，动作 $A_t = y_t$，transition 把 token append 到 prefix，EOS/上限终止，奖励可在 terminal 或逐 token 给出。由乘积分解直接取 log：

$$\log \frac{\pi_\eta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} = \sum_t [\log \pi_\eta(y_t \mid S_t) - \log \pi_{\text{ref}}(y_t \mid S_t)].$$

support 条件保证 log-ratio 有限。

(b) 环境/prompt 不依赖 η 时，

$$\nabla_\eta \mathbb{E}[R] = \mathbb{E}\left[R(x, Y) \sum_t \nabla_\eta \log \pi_\eta(Y_t \mid S_t)\right].$$

若有逐 token reward，causality 可删除动作前奖励，以从 t 开始的 discounted reward-to-go 替换完整 return；须保持相对/绝对 discount convention 一致。

(c) 记 $L(Y) = \log(\pi_\eta(Y \mid x)/\pi_{\text{ref}}(Y \mid x))$。一般

$$\begin{aligned}\nabla J &= \mathbb{E}\left[(R - \beta L)\nabla \log \pi_\eta(Y \mid x) - \beta \nabla L\right] \\ &= \mathbb{E}\left[(R - \beta L) \sum_t \nabla \log \pi_\eta(Y_t \mid S_t)\right],\end{aligned}$$

因为 reference frozen 且 $\mathbb{E}_{Y \sim \pi_\eta} \nabla \log \pi_\eta(Y \mid x) = \nabla \sum_Y \pi_\eta(Y \mid x) = 0$。故把 sampled regularized return stop-gradient 后乘 trajectory score 给出无偏形式。

(d) Baseline 可依赖 state/prefix，但在给定 state 后不得依赖当前 sampled action；这样 $\mathbb{E}_{a \sim \pi}[\nabla \log \pi(a \mid s) b(s)] = 0$。

Rubric: MDP 与 log-ratio 3 分；trajectory score/reward-to-go 3 分；显式梯度、零期望与最终式 3 分；baseline 条件 1 分。

C. 人工智能中的优化方法：参考解答 [33 points]

(1) [5 points]

$$\nabla F_\lambda(x) = \nabla h(x) + \lambda x.$$

因此梯度 Lipschitz 常数为 $L_0 + \lambda$。凸函数 h 与 λ-strongly convex 的二次项之和为 λ-strongly convex。又因无约束最优点满足 $\nabla h(x^\star) = 0$，凸性给出 $h(x) \geq h(x^\star)$，故

$$F_\lambda(x) \geq h(x^\star) + \frac{\lambda}{2}\|x\|^2$$

是 coercive，最小点存在；强凸性保证唯一。条件数和一阶条件为

$$\kappa_\lambda = \frac{L_0 + \lambda}{\lambda}, \qquad \nabla h(x_\lambda) + \lambda x_\lambda = 0.$$

(2) [5 points] $x^\star$ 最小化 h，所以左侧非负。另一方面，

$$h(x_\lambda) + \frac{\lambda}{2}\|x_\lambda\|^2 \leq h(x^\star) + \frac{\lambda}{2}\|x^\star\|^2.$$

移项并丢弃 $\|x_\lambda\|^2$：

$$h(x_\lambda) - h(x^\star) \leq \frac{\lambda}{2}(\|x^\star\|^2 - \|x_\lambda\|^2) \leq \frac{\lambda R^2}{2}.$$

(3) [7 points] 令 $M = L_0 + \lambda$。步长为 $1/M$ 时，descent lemma 给出

$$F_\lambda(x^{k+1}) \leq F_\lambda(x^k) - \frac{1}{2M}\|\nabla F_\lambda(x^k)\|^2.$$

λ-strong convexity 蕴含

$$\|\nabla F_\lambda(x)\|^2 \geq 2\lambda[F_\lambda(x) - F_\lambda(x_\lambda)].$$

代入得到

$$F_\lambda(x^{k+1}) - F_\lambda(x_\lambda) \leq \left(1 - \frac{\lambda}{M}\right)[F_\lambda(x^k) - F_\lambda(x_\lambda)].$$

递归 K 次即得。smoothness 提供稳定下降，strong convexity 把梯度平方下界为 gap；两者共同产生几何而非次线性收缩。

(4) [6 points] 精确拆分

$$h(x^K) - h(x^\star) = F_\lambda(x^K) - F_\lambda(x_\lambda) + [F_\lambda(x_\lambda) - h(x^\star)] - \frac{\lambda}{2}\|x^K\|^2.$$

最后一项非正，且

$$F_\lambda(x_\lambda) \leq F_\lambda(x^\star) \leq h(x^\star) + \frac{\lambda R^2}{2}.$$

合并即得题设式。第一项是新目标尚未优化完的 optimization error，第二项是改变目标造成的 bias 上界。

(5) [6 points] 取 $\lambda = \varepsilon/R^2$ 后，bias 至多 $\varepsilon/2$。令 $a = \lambda/(L_0 + \lambda)$，并用 $(1-a)^K \leq e^{-aK}$。若 $\Delta_\lambda \leq \varepsilon/2$，可取 $K = 0$。否则充分取

$$K \geq \left\lceil \frac{L_0 + \lambda}{\lambda} \log \frac{2\Delta_\lambda}{\varepsilon} \right\rceil.$$

此时 optimization error 也不超过 $\varepsilon/2$。增大 λ 会降低 κ_λ、加快新目标优化，却线性增大 bias 上界 $\lambda R^2/2$。

(6) [4 points]

$$x_\lambda = (A^\top A + \lambda I)^{-1} A^\top b.$$

若极端奇异值为 $\sigma_{\max}, \sigma_{\min}$，则 Hessian 条件数为

$$\frac{\sigma_{\max}^2 + \lambda}{\sigma_{\min}^2 + \lambda}.$$

即使 $\sigma_{\min} = 0$，加入 λI 后最小特征值为 $\lambda > 0$，所以正则化问题强凸且解唯一。

Rubric: (1) smooth/strong 2 分、existence/uniqueness 1 分、条件数/最优性 2 分；(2) comparison 3 分、两侧与常数 2 分；(3) descent 2 分、strong-convex gap 2 分、递归与解释 3 分；(4) exact decomposition 3 分、上界 2 分、解释 1 分；(5) bias split 1 分、两个 K 情形 3 分、tradeoff 2 分；(6) ridge 解 1 分、条件数 2 分、秩亏解释 1 分。

D. 自然语言处理：参考解答 [33 points]

1. Orthogonal cross-lingual alignment [8 分]

利用 $W^\top W = I$ 展开：

$$\|WX - Y\|_F^2 = \operatorname{tr}(X^\top X) + \operatorname{tr}(Y^\top Y) - 2\operatorname{tr}(W^\top Y X^\top).$$

故等价于最大化 $\operatorname{tr}(W^\top U\Sigma V^\top)$。令 $Q = U^\top W V$，则 Q orthogonal，且

$$\operatorname{tr}(W^\top U\Sigma V^\top) = \operatorname{tr}(Q^\top \Sigma) \le \sum_i \sigma_i.$$

取 $Q = I$ 达到上界，因此

$$\boxed{W^* = UV^\top}.$$

若 $YX^\top$ rank deficient，则 zero-singular-value subspace 上的 orthogonal action 不影响目标，最优解在该 subspace 可不唯一。若另要求 $\det W = 1$，reflection correction 才需要额外处理；本题没有该限制。

Orthogonal map 保持 source space 内的 norms、inner products 与 Euclidean distances，避免为了拟合 anchors 任意 stretch。失败原因包括两语言 embedding spaces 非近似 isomorphic、polysemy/context 对应不同、anchor dictionary 有系统偏差，或需要 nonlinear/context-dependent alignment。

评分要点 trace expansion 2 分；SVD/change of variables 与 upper bound 2 分；W^* 与 rank-deficient 说明 1 分；preserved geometry 1 分，具体 distant-language failure 2 分。

2. Estimating a scaling exponent [7 分]

令 excess $e(N) = \mathcal{L}(N) - \mathcal{L}_\infty$，则

$$\boxed{\log e(N) = \log A - \alpha \log N}.$$

因此以 $\log N$ 为横轴、$\log e$ 为纵轴时 slope 为 $-\alpha$。

由

$$\frac{e(4N)}{e(N)} = 4^{-\alpha} = \frac{0.10}{0.20} = \frac{1}{2}$$

得

$$\boxed{\alpha = \frac{\log 2}{\log 4} = \frac{1}{2}}.$$

从 $4N$ 到 $16N$ 又扩大四倍，excess 再乘 $4^{-1/2} = 1/2$，故

$$\boxed{e(16N) = 0.05}.$$

真实观测的是 $\mathcal{L} - \mathcal{L}_\infty$；若 floor 估错，取 log 前的 vertical differences 被 nonlinearly 扭曲，slope 也会偏。通常需要三个以上 scales 联合拟合 $\mathcal{L}_\infty, A, \alpha$ 并给 confidence interval。外推可因 data quality/重复、optimization error、architecture 或 tokenizer 改变、compute bottleneck、finite-data saturation 等失效，任答一项具体原因。

评分要点 log-linear relation/slope 2 分；α 2 分、prediction 1 分；floor effect 1 分、extrapolation failure 1 分。

3. Length bias in sequence decoding [7 分]

Raw probabilities 为

$$p(A) = 0.7^2 = 0.49, \qquad p(B) = 0.8^4 = 0.4096,$$

故 raw log probability 选择 A。Average log-scores 为

$$\frac{1}{2}\log p(A) = \log 0.7 \approx -0.3567, \qquad \frac{1}{4}\log p(B) = \log 0.8 \approx -0.2231,$$

故 average log probability 选择 B。

每个 conditional probability 至多为 1，继续生成通常加入一个非正 log term，因此 raw sum 对长序列有累积惩罚；EOS calibration 也直接影响长度。用 $T^{-1}\log p(y)$ 只是改变 ranking score，它不保留原 joint $p(y)$ 的概率比，也没有提供跨所有 variable-length sequences 的新 global normalizer；不能把它直接称为同一个 normalized joint likelihood。

Beam 只保留每一步 top-b prefixes；某个当前分数较低但以后完成概率较高的 prefix 可被永久剪掉，所以有限 b 不保证 global MAP。增大 b 更接近所选 decoding score 的 optimum，却可能更强化 model likelihood 的 length/repetition bias；而 human quality 还包含 factuality、diversity 与 task utility，未必与该 score 对齐。

评分要点 raw probabilities/ranking 1.5 分、average scores/ranking 1.5 分；bias 1 分、非同一 normalized joint 1 分；beam pruning 1 分、objective–quality mismatch 1 分。

4. Multilingual, privacy-preserving RAG [11 分]

Representative design：离线按 original-language sentence/chunk 建 immutable versioned records，保存 language、ACL、owner、retention、source offsets 与 deletion status；建立 per-ACL shard 或支持 trusted bitset prefilter 的 multilingual dense ANN，同时保留 per-language BM25。在线先 authenticate user，将 ACL 用作 candidate-domain hard filter，再对 original query 与受控 translation/query expansion 做 hybrid retrieval。Cross-lingual reranker 读取 query、original chunk 与可选 translation；generator 可用用户语言回答，但 citation 永远指向 authorized original span/version。Verifier 检查 claim entailment、language、ACL 与 version freshness。

令 $\mathcal{A}(u)$ 为用户 u 可访问 chunks，$k_u = \min\{k, |\mathcal{A}(u)|\}$，可定义

$$\boxed{R_k(q,u) = \arg\max_{S\subseteq\mathcal{A}(u),\ |S|=k_u} \sum_{c\in S} s(q,c).}$$

Authorization 是 feasible set 的 hard constraint，而非 score penalty。若先在 global ANN 取 top-k 再过滤，unauthorized IDs/scores 可能进入不可信 logs、timing 或 citation path；同时它们占掉 candidate slots，过滤后不足 k，使 authorized recall 降低。应采用 ACL-aware shards/prefilter，或在可信边界内 oversample 并补检。

Deletion/permission event 应生成 tombstone/version update，原子更新 dense/sparse indices 与 ACL filter，invalidate query/answer/embedding/KV caches，并使旧 citation endpoint 不再返回正

文；后台 compaction 后用 canary queries 验证不可检索。若 evidence 只在 generator 弱支持的语言中，先在受控环境翻译，要求 back-translation 或 multilingual entailment agreement；confidence 低则返回 authorized original citation 和“无法可靠翻译”的 abstention，或路由有权限的 human translator，不能凭常识补写。

Quality 按 language 报 Recall@k、answer correctness、claim-level citation precision/recall、translation adequacy 与 calibration/risk–coverage；privacy 报 unauthorized retrieval/citation rate（目标 0）、canary membership-inference AUC 与 timing gap；deletion 报 time-to-unretrievable、cache purge success、deleted- source citation rate。应同时报告 macro across languages，避免高资源语言掩盖失败。

评分要点 versioned multilingual offline/online flow 4 分；hard objective 1 分、post-filter 的 security+recall 1 分；deletion propagation 1.5 分、translation/abstention 1.5 分；quality/evidence 1 分、privacy/deletion quantitative metrics 1 分。其他满足 authorization、original citation 与 deletion 的方案同等给分。