QIUZHEN QUALIFY EXAM FOR ARTIFICIAL INTELLIGENCE
SPRING 2026

Instructions

- *This exam consists of four sections, each with a total of 33 points. You are required to choose three out of the four sections and answer the questions in the selected sections. You will earn an additional point for writing your name and student ID number on your answer sheet. The maximum possible score you can achieve is 100.*
- *Please clearly indicate your section choices at the very beginning of your answer sheet. If you answer questions in more than three sections, only the first three sections will be graded.*

A. Machine Learning Theory

Part I: Warm-Up Questions 3pts — This section contains multiple-choice questions. Please select **only one** answer for each question. No justification is required.

MQ1. [0.5pts] In the agnostic PAC learning setting, what quantity does a learning algorithm aim to compete with?
(a) The Bayes optimal classifier.
(b) The hypothesis with zero training error.
(c) The best hypothesis in the class $\mathcal{H}$.
(d) The hypothesis minimising validation error.

MQ2. [0.5 pts] Let $\mathcal{H}$ be a binary hypothesis class with VC dimension $d < \infty$. Which statement about its growth function $\tau_{\mathcal{H}}(m)$ is necessarily true?
(a) $\tau_{\mathcal{H}}(m) = 2^m$ for all m.
(b) $\tau_{\mathcal{H}}(m) = O(m^d)$ for all m.
(c) $\tau_{\mathcal{H}}(m) \leqslant \sum_{i=0}^{d} \binom{m}{i}$ for all m.
(d) $\tau_{\mathcal{H}}(m)$ is constant for $m > d$.

MQ3. [0.5 pts] Which statement about ReLU neural networks is correct from a learning-theoretic perspective?
(a) ReLU networks represent smooth functions on $\mathbb{R}^d$.
(b) Increasing depth always decreases the VC dimension.
(c) ReLU networks compute piecewise linear functions whose complexity depends on depth and width.

(d) Universal approximation of ReLU networks implies PAC learnability.

MQ4. [0.5pts] Which statement about empirical risk minimisation (ERM) is correct?

(a) ERM is guaranteed to be consistent for any hypothesis class.
(b) ERM always finds a hypothesis with minimum true risk.
(c) If $\mathcal{H}$ has finite VC dimension, ERM is PAC learnable in the realisable setting.
(d) ERM is PAC learnable only if $\mathcal{H}$ is finite.

MQ5. [0.5pts] Which statement is a correct consequence of the No Free Lunch theorem?

(a) Restricting the hypothesis class always improves generalization.
(b) There exists a universally optimal learning algorithm.
(c) Inductive bias is necessary to obtain non-trivial learning guarantees.
(d) Random guessing is optimal for all supervised learning problems.

MQ6. [0.5pts] Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function. Which of the following statements is *correct*?

(a) K is a valid kernel if and only if $K(x, x) \geqslant 0$ for all $x \in \mathcal{X}$.
(b) K is a valid kernel if and only if there exists a finite-dimensional feature map ψ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle$.
(c) If K is positive semidefinite, then there exists a (possibly infinite-dimensional) Hilbert space $\mathcal{H}$ and a feature map $\psi : \mathcal{X} \to \mathcal{H}$ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}}$.
(d) Every kernel corresponds to a unique feature map.

Part II: Theoretical Exercises 30pts — This is the main part of the exam. Provide **detailed solutions and reasoning** for each question. Full marks are awarded only for complete and well-explained answers.

(1) [6 pts.] Let $\mathcal{H}$ be a hypothesis class of binary classifiers $h : \mathcal{X} \to \{0, 1\}$. Let D be an unknown distribution over $\mathcal{X}$, and let $f \in \mathcal{H}$ be the target hypothesis.
For a fixed $h \in \mathcal{H}$, define the empirical loss on a sample $S = \{x_1, \ldots, x_m\} \sim D^m$ by

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[h(x_i) \neq f(x_i)] .$$

(a) Let $p := L_{D,f}(h)$. Show that

$$\mathbb{E}_{S \sim D^m} \Big[\big(L_S(h) - p \big)^2 \Big] = \frac{p(1-p)}{m}.$$

(b) Deduce that

$$\mathbb{E}\Big[\big(L_S(h) - p \big)^2 \Big] \leqslant \frac{1}{4m},$$

and therefore it decreases at rate $O(1/m)$.

(2) [4 pts.] Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and λ-strongly convex. Prove that f admits a unique global minimiser.

(3) [8 pts.]Let $\mathcal{X} = \mathbb{R}$ and consider the hypothesis class

$$\mathcal{H} = \left\{ h_{a,b,c}(x) = \mathbf{1}_{[a,b]}(x) \vee \mathbf{1}_{[c,\infty)}(x) \;\middle|\; a \leqslant b < c,\ a,b,c \in \mathbb{R} \right\}.$$

where $\vee$ denotes the logical OR (equivalently, the maximum of the two values).

(a) Let $x_1, \ldots, x_n$ be n distinct points in $\mathbb{R}$. Give an upper bound on the growth function $s(\mathcal{H}, n)$.

(b) Determine the VC dimension of $\mathcal{H}$.

(4) [7 pts.] Consider a set S of examples in $\mathbb{R}^n \times [k]$ for which there exist vectors $\mu_1, \ldots, \mu_k \in \mathbb{R}^n$ such that every example $(x, y) \in S$ falls within a ball centered at μ_y whose radius is $r \geqslant 1$. Assume also that for every $i \neq j$, $\|\mu_i - \mu_j\| \geqslant 4r$.
Consider concatenating each instance by the constant 1 and then applying the multivector construction, namely:

$$\Psi(x,y) = [\ \underbrace{0, \ldots, 0}_{(y-1)(n+1)}, \underbrace{x_1, \ldots, x_n, 1}_{n+1}, \underbrace{0, \ldots, 0}_{(k-y)(n+1)}\ ] \in \mathbb{R}^{k(n+1)}.$$

Note that the 1 is referring to the bias term. It allows the model to shift the decision boundary away from the origin. Show that there exists a vector $w \in \mathbb{R}^{k(n+1)}$ such that $\ell(w, (x, y)) = 0$ for every $(x, y) \in S$.

Hint. Observe that for every example $(x, y) \in S$ we can write $x = \mu_y + v$ for some $\|v\| \leqslant r$. Now take $\mathbf{w} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ where:

$$\mathbf{w}_i = \left[\boldsymbol{\mu}_i,\ -\frac{\|\boldsymbol{\mu}_i\|^2}{2} \right].$$

(5) [5 pts.] Let $d \in \mathbb{N}$. Consider two feedforward neural networks with the same input dimension d and scalar output. Assume that the first network Φ_{ReLU} uses the ReLU activation function in all hidden layers, while the second network Φ_σ uses the sigmoid activation function. Suppose that the two networks represent the same function on $\mathbb{R}^d$, that is:

$$\Phi_{\text{ReLU}}(x) = \Phi_\sigma(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Show that the function represented by Φ_{ReLU} (and hence by Φ_σ) must be constant.

B. Deep Learning and Reinforcement Learning

(1) [10 pts.]

(a) [2 pts.] Consider a multi-layer perceptron (MLP) with layerwise relation

$$x_\ell = f_\ell(x_{\ell-1}, \theta_\ell), \qquad \ell = 1, \ldots, L,$$

where θ_ℓ denotes the parameters of layer ℓ, and a loss $\mathcal{J} = \mathcal{L}(x_L, y)$.

(i) Using chain rule, derive the expression for $\frac{\partial \mathcal{J}}{\partial \theta_\ell}$.

(ii) State briefly why computing all gradients $\left\{ \dfrac{\partial \mathcal{J}}{\partial \theta_\ell} \right\}_{\ell=1}^{L}$ via backpropagation has the same order of computational complexity as one forward evaluation (up to a constant factor).

(b) [2 pts.] Explain mathematically why the **sigmoid** activation

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

often leads to the **vanishing gradient problem** in deep networks. Contrast this with the **ReLU** activation and explain how ReLU mitigates this issue during backpropagation.

(c) [2 pts.] Write the update rules for **stochastic gradient descent (SGD) with momentum**. Define the momentum term v_{k+1} and the parameter update x_{k+1}. Explain how the momentum term improves optimization compared to standard SGD, particularly in loss landscapes with oscillations.

(d) [2 pts.] The scaled dot-product attention in Transformers is defined as

$$W = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} \right).$$

Explain the purpose of the scaling factor $1/\sqrt{d_k}$. What happens to the softmax outputs and their gradients if this scaling is removed when d_k is large?

(e) [2 pts.] Compare the computational structure of **recurrent neural networks (RNNs)** and **Transformers**. Explain why Transformers are typically more computationally efficient than RNNs in modern hardware implementations.

(2) [15 pts.] Let π_0 be a simple base distribution on $\mathbb{R}^d$ (e.g., $\mathcal{N}(0, I)$) and let π_1 be a data distribution. We seek a time-dependent vector field $v_t(x)$ such that the ODE

$$\frac{d}{dt} x_t = v_t(x_t), \qquad x_0 \sim \pi_0,$$

transports π_0 to π_1 at time $t = 1$.

(a) [2 pts.] Assume a prescribed coupling $(x_0, x_1) \sim \gamma$ with marginals $x_0 \sim \pi_0$, $x_1 \sim \pi_1$, and define the interpolation

$$x_t = (1 - t)x_0 + t x_1.$$

Compute the *conditional velocity*

$$v_t(x_t \,|\, x_0, x_1) := \frac{d}{dt} x_t.$$

(b) [3 pts.] Define the population velocity field

$$v_t(x) = \mathbb{E}[v_t(x_t \,|\, x_0, x_1) \mid x_t = x].$$

Show that if particles evolve according to

$$\frac{d}{dt} x_t = v_t(x_t),$$

then the density π_t of x_t satisfies the continuity equation

$$\partial_t \pi_t + \nabla\cdot (\pi_t v_t) = 0.$$

Conclude that this ODE transports π_0 to π_1.

(c) [5 pts.] Let $v_\theta(t, x)$ be a parametric model. Propose a squared-loss regression objective that uses training pairs (t, x_t) and the conditional velocity $v_t(x_t \mid x_0, x_1)$ to learn $v_\theta(t, x)$. Prove that the minimizer of the loss function recovers the desired marginal velocity $v_t(x)$.

(d) [5 pts.] To introduce stochasticity, consider the SDE

$$dx_t = \Big(v_t(x_t) + \beta_t \nabla_x \log \pi_t(x_t)\Big)dt + \sqrt{2\beta_t}\, dW_t.$$

Derive the Fokker–Planck equation of this SDE and show that this SDE has the *same marginal distributions* $\{\pi_t\}_{t\in[0,1]}$ as the deterministic ODE.

In practice $\nabla_x \log \pi_t(x)$ is unknown. Explain how this term can be approximated using *score matching.*

(3) [8 pts.] Consider a discounted Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\gamma \in (0, 1)$ is the discount factor, $P(s'|s, a)$ is the transition kernel, and $r(s, a)$ is the reward.

(a) [2 pts.] The value function under a policy $\pi(a|s)$ is defined as

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right].$$

Derive the Bellman equation satisfied by $V^\pi(s)$.

(b) [2 pts.] Write down the **policy iteration algorithm**. Clearly specify:

- the policy evaluation step, and
- the policy improvement step.

(c) [2 pts.] In many practical problems the transition model P is unknown. Describe how the evaluation and improvement steps are adapted using an **actor–critic** method under this *model-free* setting.

(d) [2 pts.] Reinforcement Learning from Human Feedback (RLHF) aligns a policy using human preference data of the form (x, y^+, y^-), meaning y^+ is preferred to y^- for prompt x.

Write down **one RLHF algorithm** by specifying the objective used to update the policy $\pi_\theta(y|x)$.

Hint: You may choose either of the following approaches:

(a) **Reward-model-based RLHF:** first learn a reward model $r_\psi(x, y)$ from preference data, then optimize the policy using a reinforcement learning objective (often with a KL regularization toward a reference policy).

(b) **Direct Preference Optimization (DPO)-style methods:** update the policy directly from preference comparisons without explicitly learning a reward model, by encouraging higher likelihood of preferred outputs relative to dispreferred ones.

Define any notation you introduce.

C. Optimization Methods in Artificial Intelligence

Consider the composite optimization problem

$$\min_{x\in\mathbb{R}^d} F(x) := f(x) + R(x),$$

where f is differentiable L-smooth and μ-strongly convex ($\mu > 0$), and $R : \mathbb{R}^d \to (-\infty, +\infty]$ is proper, closed, and convex. Let $g(x, \xi) = \nabla f(x) + \xi$ be an unbiased stochastic gradient estimator satisfying $\mathbb{E}[\xi] = 0$ and $\mathbb{E}\|\xi\|^2 \leqslant \sigma^2$. Consider the proximal SGD iteration

$$x^{k+1} = \operatorname{prox}_{\gamma R}\left(x^k - \gamma g^k\right), \qquad g^k := g(x^k, \xi^k) = \nabla f(x^k) + \xi^k.$$

Define the averaged iterate

$$\bar{x}^k := \frac{1}{k}\sum_{t=1}^{k} x^t.$$

(1) [2 pts.] *Basic definitions.* Give the definitions of L-smoothness and μ-strong convexity.

(2) [3 pts.] *Optimality condition of the proximal operator.* Show that for any $u \in \mathbb{R}^d$ and $\gamma > 0$,

$$x^+ = \operatorname{prox}_{\gamma R}(u) \quad \Longleftrightarrow \quad \frac{1}{\gamma}(u - x^+) \in \partial R(x^+).$$

(3) [3 pts.] *Uniqueness and proximal fixed point.* Show that F admits a unique minimizer $x^\star$ and that

$$x^\star = \operatorname{prox}_{\gamma R}(x^\star - \gamma \nabla f(x^\star)).$$

(4) [2 pts.] *Monotonicity of the subdifferential.* Show that the subdifferential mapping ∂R is monotone: for any x, y and any $r_x \in \partial R(x)$, $r_y \in \partial R(y)$,

$$\langle r_x - r_y,\, x - y\rangle \geqslant 0.$$

(5) [3 pts.] *Nonexpansiveness of* prox. Show that $\operatorname{prox}_{\gamma R}$ is nonexpansive, i.e.,

$$\|\operatorname{prox}_{\gamma R}(u) - \operatorname{prox}_{\gamma R}(v)\| \leqslant \|u - v\|, \qquad \forall u, v.$$

(*Hint:* use monotonicity of ∂R.)

(6) [3 pts.] *One-step inequality.* Show that

$$\|x^{k+1} - x^\star\|^2 \leqslant \|x^k - x^\star - \gamma(g^k - \nabla f(x^\star))\|^2.$$

(7) [5 pts.] *Conditional expectation and variance decomposition.* Write $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid x^k]$ for the conditional expectation given the current iterate x^k. Prove that

$$\mathbb{E}_k\|x^{k+1} - x^\star\|^2 \leqslant \|x^k - x^\star - \gamma(\nabla f(x^k) - \nabla f(x^\star))\|^2 + \gamma^2\sigma^2.$$

(8) [4 pts.] *Linear contraction with a noise floor.* Using smoothness and strong convexity of f, show that for $\gamma = \frac{2}{L+\mu}$ one has

$$\mathbb{E}\|x^{k+1} - x^\star\|^2 \leqslant (1-\rho)\,\mathbb{E}\|x^k - x^\star\|^2 + \gamma^2\sigma^2, \qquad \rho := \frac{4\mu L}{(L+\mu)^2} \in (0,1).$$

Explain the origin and meaning of the *noise floor.*
Hint: You may use the following two inequalities for an L-smooth and μ-strongly convex function f:

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geqslant \mu\|x-y\|^2, \qquad \langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geqslant \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

(9) [3 pts.] *Averaging reduces variance.* Write the iterate error as $e^t := x^t - \mathbb{E}[x^t]$ and define the averaged error

$$\bar{e}^k := \bar{x}^k - \mathbb{E}[\bar{x}^k] = \frac{1}{k}\sum_{t=1}^{k} e^t.$$

Assume (as a simplified model):

- the errors $\{e^t\}_{t\geqslant 1}$ are uncorrelated, i.e., $\mathbb{E}\langle e^s, e^t\rangle = 0$ for $s \neq t$;
- the second moments are uniformly bounded: $\mathbb{E}\|e^t\|^2 \leqslant V$ for all t.

Show that

$$\mathbb{E}\|\bar{e}^k\|^2 \leqslant \frac{V}{k}.$$

Explain briefly why $\bar{x}^k$ typically has smaller variance than the last iterate x^k.

(10) [5 pts.] *Convergence rate of averaging.* For this question, assume $R \equiv 0$ (so $F = f$).

(a) [3 pts.] *Deriving a one-step bound from smoothness.* Starting from the inequality proved earlier in Question 7, show that

$$\mathbb{E}_k\|x^{t+1} - x^\star\|^2 \leqslant \|x^t - x^\star\|^2 - 2\gamma_t(1 - L\gamma_t)\,\mathbb{E}_k\big(f(x^t) - f^\star\big) + \gamma_t^2\sigma^2.$$

In particular, if $\gamma_t \leqslant \frac{1}{2L}$, deduce the simplified bound

$$\mathbb{E}_k\|x^{t+1} - x^\star\|^2 \leqslant \|x^t - x^\star\|^2 - \gamma_t\,\mathbb{E}_k\big(f(x^t) - f^\star\big) + \gamma_t^2\sigma^2.$$

(*Hint:* For convex L-smooth f, we have $f\big(x - \frac{1}{L}\nabla f(x)\big) \leqslant f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$.)

(b) [2 pts.] Choose $\gamma_t = \frac{1}{\mu t}$ and show that

$$\mathbb{E}\big[f(\bar{x}^k) - f^\star\big] \leqslant \mathcal{O}\bigg(\frac{\log k}{k}\bigg).$$

(*Hint:* Using the fact that $1 + \frac{1}{2} + \cdots + \frac{1}{k} \sim \log k$.)

D. Natural Language Processing

(1) [6 pts.] You are given an encoder $f_\theta(\cdot)$ that maps an input sentence x to a unit-norm embedding $z = f_\theta(x) \in \mathbb{R}^d$ with $\|z\|_2 = 1$. For each anchor x, you sample one positive view x^+ and n negatives $\{x_i^-\}_{i=1}^n$ from the minibatch. Denote $z^+ = f_\theta(x^+)$ and $z_i^- = f_\theta(x_i^-)$. Consider the InfoNCE loss with temperature $\tau > 0$:

$$\mathcal{L} = -\log \frac{\exp(\langle z, z^+\rangle/\tau)}{\exp(\langle z, z^+\rangle/\tau) + \sum_{i=1}^n \exp(\langle z, z_i^-\rangle/\tau)}.$$

(a) [3 pts.] (Gradient) Compute $\nabla_z \mathcal{L}$. Interpret your result as "pulling" z toward z^+ and "pushing" z away from negatives on the unit hypersphere (a tangent-space geometric interpretation is sufficient; you do *not* need to explicitly compute the projection).

(b) [3 pts.] (High-dimensional heuristic) Assume $d \gg 1$ and that negatives are isotropic on the unit sphere. Give a heuristic for the *typical scale* of $\langle z, z_i^-\rangle$ (e.g., its variance / concentration around 0), and for the *largest* similarity among n negatives, $\max_{1\leqslant i\leqslant n}\langle z, z_i^-\rangle$. Then, explain why such false negatives (some sampled negative x_i^- semantically equivalent or from the same latent class as the anchor x so with embedding close to z) can significantly hurt contrastive learning.

(2) [6 pts.] You are given a corpus of customer-support tickets from an e-commerce platform. Each ticket d is a short document with tokens $w_{dn} \in \{1, \ldots, V\}$ (after preprocessing). In addition, each ticket has an observed priority label $y_d \in \{0, 1\}$ (e.g., normal vs. urgent). The goal is to discover topics that both explain the text and help predict y_d.

Consider a supervised topic model with K topics and the following random variables:

- document-topic proportions θ_d and topic-word distributions ϕ_k;
- token-level topic assignments z_{dn} and observed words w_{dn};
- a document label $y_d \in \{0, 1\}$.

Assume Dirichlet priors $\theta_d \sim \mathrm{Dir}(\alpha)$ and $\phi_k \sim \mathrm{Dir}(\eta)$. For each document d of length N_d, define the empirical topic-frequency vector $\bar{z}_d$ where $\bar{z}_{dk} = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbb{I}[z_{dn} = k]$.

(a) [2 pts.] (Modeling) Briefly explain why adding y_d changes the learned topics compared with unsupervised LDA. Give one failure mode of vanilla LDA for this application that supervision can mitigate.

(b) [2 pts.] (Label model) Propose a probabilistic model for $p(y_d \mid \bar{z}_d, \beta)$ with parameters $\beta \in \mathbb{R}^K$. You may choose logistic regression; if so, define $\sigma(u)$ and write $p(y_d \mid \bar{z}_d, \beta)$ explicitly. Briefly interpret β.

(c) [2 pts.] (Sampling process) *Conditioned on* (θ_d, Φ, β) for a fixed document d, specify the sampling steps that generate $(z_{d1:N_d}, w_{d1:N_d}, y_d)$, and name the distribution used at each step. Then, write the full joint probability (in factorized form; you do not need to expand normalization constants)

$$p(\mathbf{w}, \mathbf{z}, \Theta, \Phi, \mathbf{y} \mid \alpha, \eta, \beta),$$

where $\Theta = \{\theta_d\}_{d=1}^D$ and $\Phi = \{\phi_k\}_{k=1}^K$, and $\mathbf{w}, \mathbf{z}, \mathbf{y}$ denote all tokens/topics/labels. Your answer should make explicit the product structure over documents, topics, and token positions, and include the label term $p(y_d \mid \bar{z}_d, \beta)$.

(3) [7 pts.] Standard self-attention costs $\mathcal{O}(L^2)$ in sequence length L. Design a Transformer variant that can handle $L \gg 10^4$ while preserving both local and long-range dependencies. Choose *one* mechanism: (i) block-sparse attention, (ii) sliding window + global tokens, (iii) low-rank/linear attention, or (iv) recurrent memory. Your design must include a precise definition of the attention pattern/computation (e.g., an attention mask M_{ij} or neighbor set $\mathcal{N}(i)$ for sparse patterns, a feature map $\phi(\cdot)$ for linear attention, or explicit read/write equations for memory).

(a) [3 pts.] (Mechanism) Propose your chosen mechanism and define its computation/pattern precisely. Introduce any needed hyperparameters (e.g., window size w, number of global tokens g, block size b, memory length m) and specify exactly how information is aggregated across positions. Keep it brief.

(b) [2 pts.] (Complexity) Analyze the *per-layer* time and memory complexity of the main component (dominant terms) in terms of sequence length L, hidden size d, and your hyperparameters. You may ignore the feed-forward sublayer cost. Keep it brief.

(c) [2 pts.] (Expressivity for long-range interactions) Give one argument why long-range interactions remain expressible under your mechanism. Your argument should be based on, e.g., attention-graph connectivity / diameter (how information from position i can reach j), or information propagation across layers via global tokens / blocks / memory.

(4) [7 pts.] You are building an information-extraction system on news articles. A pipeline produces uncertain local predictions: (i) mention-to-entity linking candidates with confidence signals, and (ii) relation candidates between entities with confidence signals. You want a global model that enforces consistency and integrates evidence.

An MLN defines a conditional distribution

$$P(X \mid E) = \frac{1}{Z(E)} \exp\Big(\sum_{j=1}^{J} w_j \, n_j(X, E)\Big),$$

where E denotes observed evidence, X denotes the set of unobserved atoms to be inferred, and $n_j(X, E)$ is the number of satisfied groundings of formula j. We use the following predicates:

- *Unknown (to infer):* $\mathrm{Link}(m, e)$, meaning mention m is linked to entity e; and $\mathrm{Rel}(e_1, e_2, r)$, meaning relation type r holds between entities (e_1, e_2).
- *Evidence (observed):* $\mathrm{Cand}(m, e)$, indicating that e is a candidate entity for mention m; $\mathrm{Type}(e, t)$, indicating that entity e has type t (e.g., Person, City, Org); $\mathrm{HighLink}(m, e)$, indicating strong local support for $\mathrm{Link}(m, e)$; and $\mathrm{HighRel}(m_1, m_2, r)$, indicating strong local support for relation type r between mentions (m_1, m_2).

Answer following questions:

(a) [2 pts.] (Modeling rules with weights) Using the predicates above, write MLN formulas that: (i) propagate HighLink and HighRel into the unknown variables Link through Rel. (ii) implement the constraint of *at most one* linked entity per mention, and (iii) enforce relation–type consistency using $\mathrm{Type}(e, t)$ (e.g., argument-type constraints). Briefly discuss how you would set (relative) weights for each rule, and which constraints you would treat as hard vs. soft.

(b) [2 pts.] (Evidence construction) In real systems, $\mathrm{HighLink}(m, e)$ and $\mathrm{HighRel}(m_1, m_2, r)$ are produced by upstream machine learning models. Describe *concrete* modeling approaches (architectures and objectives) to obtain these signals. Your answer should mention what the model inputs are, what it outputs (score/probability), and how HighLink/HighRel is derived from that output (e.g., thresholding, top-k with margin). Keep it brief.

(c) [3 pts.] (Name disambiguation) Name ambiguity is severe for people in the same domain. Suppose two candidate person entities e_{old} and e_{new} share the same surface name, but their birth years differ by 500 years. You want to leverage (i) relation evidence among people in the document and (ii) KB birth-year attributes to impose an implicit "same-era" constraint.Please complete and extend the MLN by introducing additional predicates and rules. Then provide the additional key formulas and their intended weight behaviors, showing how relation evidence (HighRel) induces contemporaneity (possibly conditioned on r) and how birth-year attributes penalize impossible contemporaneity. Finally, provide a minimal, reasonable example context (mentions, candidate entities, and one or two relation types) explaining how the joint inference prefers linking the ambiguous mention to e_{old} or e_{new}. Use a minimal example only.

(5) [7 pts.] A dialog-based NLU/QA system uses RAG as an external memory. At each turn t, given dialog history H_t, the system generate a candidate memory $c_t = \mathrm{Memorize}(H_t)$ and *write* into a store $\mathcal{M}_t$ (possibly with compression/tags), then

retrieve top-k memories $R_t = \text{Retrieve}(H_t, \mathcal{M}_t, k)$, and generate an answer $\hat{y}_t = f_\theta(H_t, R_t)$. It may also write back summaries to update $\mathcal{M}_{t+1}$.

(a) [3 pts.] (Memory Gain) Let the test set be $\mathcal{D} = \{(H_t, y_t^*)\}$ and $s(\hat{y}_t, y_t^*) \in [0,1]$ be the evaluation score. Define the per-turn gain and total gain of the memory system as

$$\Delta_t = s(f_\theta(H_t, R_t), y_t^*) - s(f_\theta(H_t, \varnothing), y_t^*), \qquad \Delta = \mathbb{E}_{(H_t, y_t^*) \sim \mathcal{D}} \Delta_t.$$

Define $C_t = 1$ meaning R_t contains the key memory; otherwise $C_t = 0$ with $\Pr(C_t = 0) = \epsilon$. Let $g_c = \mathbb{E}[\Delta_t \mid C_t = 1]$ and $g_{\neg c} = \mathbb{E}[\Delta_t \mid C_t = 0]$. Derive a necessary and sufficient condition for $\Delta > 0$ in terms of $\epsilon, g_c, g_{\neg c}$. In particular, when $g_c > g_{\neg c}$, show it is equivalent to

$$\epsilon < \frac{g_c}{g_c - g_{\neg c}}.$$

Explain in 2–3 sentences why a large ϵ makes stable positive gain hard to guarantee. (3 points)

(b) [4 pts.] (Optimization) Formulate the memory mechanism as an MDP and optimize it with RL. Specify the state s_t, action a_t, and reward r_t. Your action should explicitly include both (i) *write/merge/evict* decisions under the budget constraint, and (ii) *retrieval* decisions (e.g., k, hybrid weights, query rewriting, reranking). Propose a reasonable reward that trades off answer quality and memory cost, e.g., $r_t = \Delta_t - \lambda \cdot \text{Cost}(a_t)$, and state the optimization objective.