Supplementary Materials for BayesAgent: Bayesian Agentic Reasoning Under Uncertainty via Verbalized Probabilistic Graphical Modeling

1 Proof of Theorem 1

Theorem 1 (**Global Optimum Implies Perfect ECE**). *Let $\{(\mathbf{u}_i, y_i)\}_{i=1}^n$ be the training set with features $\mathbf{u}_i \in \mathbb{R}^d$ and one–hot labels y_{ik}. For any parameter vector θ, let $g_\theta : \mathbb{R}^d \to \Delta^{K-1}$ be a function that produces class probabilities $\widehat{p}_{ik}(\theta) = g_\theta(\mathbf{u}_i)_k$. The empirical version of $\mathcal{L}\big(\boldsymbol{\pi}(\lambda)\big) \; = \; \mathcal{L}_c\big(\boldsymbol{\pi}(\lambda)\big) \; + \; \beta\, \mathcal{L}_v\big(\boldsymbol{\pi}(\lambda)\big)$ is*

$$\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\, \log \widehat{p}_{ik}(\theta) + \beta\, \frac{1}{K}\sum_{k=1}^{K} \big|\bar{\widehat{p}}_k(\theta) - \bar{y}_k\big|, \qquad \beta > 0,$$

where $\bar{\widehat{p}}_k(\theta) = \frac{1}{n}\sum_i \widehat{p}_{ik}(\theta)$, and $\bar{y}_k = \frac{1}{n}\sum_i y_{ik}$. Then a parameter vector $\theta^\star$ is a global minimiser of $\mathcal{L}$ iff

$$\widehat{p}_{ik}(\theta^\star) = \frac{1}{\sum_{i'=1}^{n} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}} \sum_{i'=1}^{n} y_{i'k} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}, \quad \text{for every } i \in \{1, \ldots, n\},\ k \in \{1, \ldots, K\},$$

where $\mathbb{I}(\mathbf{u}_i = \mathbf{u}_{i'})$ is an indicator function equal to 1 if the feature inputs $\mathbf{u}_i$ and $\mathbf{u}_{i'}$ are identical, and 0 otherwise. In that case, the class-wise expected calibration error $\mathrm{ECE}_{\text{class}}(\theta) \triangleq \frac{1}{K}\sum_k |\bar{\widehat{p}}_k(\theta) - \bar{y}_k|$ satisfies $\mathrm{ECE}_{\text{class}}(\theta^\star) = 0$.

Proof. The two summands of $\mathcal{L}$ are non-negative: the first by Gibbs' inequality, the second by the modulus. Hence $\mathcal{L}(\theta) \geq 0$ for all θ.

Sufficiency. If $\widehat{p}_{ik}(\theta^\star) = \frac{1}{\sum_{i'=1}^{n} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}} \sum_{i'=1}^{n} y_{i'k} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}$ for every pair (i, k), then the cross-entropy term vanishes and $\bar{\widehat{p}}_k(\theta) = \bar{y}_k$ for each k, forcing the calibration term to vanish as well; thus $\mathcal{L}(\theta) = 0$. No smaller value is attainable, so such a parameter vector is a global minimiser and $\mathrm{ECE}_{\text{class}}(\theta) = 0$.

Necessity. Suppose $\theta^\star$ minimises $\mathcal{L}$ yet there exists a pair (i, k) with $\widehat{p}_{ik}(\theta^\star) \neq \frac{1}{\sum_{i'=1}^{n} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}} \sum_{i'=1}^{n} y_{i'k} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}$. Then the cross-entropy term is strictly positive, implying $\mathcal{L}(\theta^\star) > 0$. However, the parameter choice described above achieves $\mathcal{L} = 0$, contradicting optimality. Therefore $\widehat{p}_{ik}(\theta^\star) = \frac{1}{\sum_{i'=1}^{n} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}} \sum_{i'=1}^{n} y_{i'k} \mathbb{I}_{\{\mathbf{u}_i = \mathbf{u}_{i'}\}}$ for all (i, k). With these equalities in place, $\bar{\widehat{p}}_k(\theta^\star) = \bar{y}_k$ for every class, so the calibration component – and hence the class-wise $\mathrm{ECE}_{\text{class}}$ – is zero. $\square$

2 Background: Probabilistic Graphical Models in Bayesian Inference

Probabilistic Graphical Models (PGMs) are powerful tools for representing uncertainty and dependencies among variables [4, 5]. We focus on *Bayesian Networks* (BNs), a directed class of PGMs whose nodes correspond to random variables and whose edges encode conditional dependencies in a directed acyclic graph (DAG). Concretely, a BN over n latent variables $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ factors their joint distribution as

$$P(\mathbf{Z}) \; = \; \prod_{i=1}^{n} P\big(Z_i \;\mid\; \mathrm{Pa}(Z_i)\big), \tag{1}$$

where $\mathrm{Pa}(Z_i)$ denotes the parent nodes of Z_i. Each term $P\big(Z_i \mid \mathrm{Pa}(Z_i)\big)$ is called a *conditional probability distribution* (CPD), and it specifies how a variable depends on its parents in the DAG.

Within the Bayesian paradigm, model parameters (i.e., of each CPD) are initially assigned with priors; as new data arrive, Bayesian inference refines these priors into posteriors, thereby capturing revised beliefs. However, designing a DAG and estimating its parameters can be challenging, especially when data are scarce or when domain expertise is limited. In this work, we overcome these constraints by leveraging Large Language Models (LLMs) to *verbalize*, discover, and perform inference in a simulated or verbalized Bayesian network without conventional data-intensive training or expert-defined structures, thus broadening the applicability of PGMs.

3 More Detailed Experiment Setup and Results

LLM Configuration We use GPT-4 [1] for PGM discovery and constructing Bayesian inference instructions for **vPGM**, while GPT-3.5-turbo-1106 and Llama3-8B-Instruct [2] (on two NVIDIA A100 GPUs) serve as our test-time engine for all prompting-based methods. Unless otherwise specified, the temperature is fixed at 0.2. We generate three candidate responses for vPGM and BayesVPGM to estimate confidence.

3.1 Dataset

ScienceQA To accommodate **BayesVPGM**, which requires a development set to optimize the hyperparameter λ, we randomly sample 3,568 data points from ScienceQA. Among these, 2,563 form the test set, while the remaining 1005 comprise the development set used to tune λ.

A-OKVQA Negative Control For our A-OKVQA-based experiment, we include 1,206 data points (both *clean* and *noisy* subsets) for testing and allocate 1,005 data points to the development set for hyperparameter tuning.

ChatCoach Since **BayesVPGM** is not applied to ChatCoach (which produces open-ended, non-categorical outputs), we use the entire dataset for evaluation. Additionally, due to ChatGPT's safety mechanisms, any prompts or responses flagged as potentially inappropriate are excluded from our reported results.

3.2 Training Detail

BayesVPGM on ScienceQA We employ L-BFGS to optimize λ through the reparameterization $\tau = 1/\lambda$, thus constraining the search space. We initialize τ to 2×10^{-5}, adopt a learning rate of 1×10^{-8}, and fix the maximum number of iterations at $1,000$.

BayesVPGM on A-OKVQA Negative Control We employ L-BFGS to optimize λ through the reparameterization $\tau = 1/\lambda$. We initialize τ to 2×10^{-12}, use a learning rate of 1×10^{-6}, and allow up to 20,000 iterations.

3.3 More Detailed Experimental Results on ScienceQA

In Table 1, We present more detailed experimental results on ScienceQA, including experiments performed on closed-sourced LLM (GPT-3.5).

3.4 Ablation Results of vPGM

For a fair comparison, all methods, including Chameleon+, vPGM, and BayesVPGM, runs under the same three-query setting, fixed latent-variable count $N = 4$, and Llama 3 as the test-time base model. Since vPGM is built on top of Chameleon+, we can interpret the variants as follows: *vPGM* = Chameleon+ & PGM; *BayesVPGM* = Chameleon+ & PGM & numerical Bayesian inference (nBI).

Method	Model	N	M	Acc. ↑	ECE ↓
CoT	GPT-3.5	–	1	83.34	19.83
Chameleon	GPT-3.5	–	1	83.93	10.63
Chameleon+	GPT-3.5	–	3	81.97	10.74
vPGM (Ours)	GPT-3.5	2	3	**84.39**	<u>2.17</u>
BayesVPGM (Ours)	GPT-3.5	2	3	**84.39**	**1.75**
CoT	Llama 3	–	1	84.63	8.96
Chameleon	Llama 3	–	1	85.29	9.62
Chameleon+	Llama 3	–	3	85.17	8.65
vPGM (Ours)	Llama 3	2	3	85.49	2.31
vPGM (Ours)	Llama 3	3	3	<u>86.38</u>	1.67
vPGM (Ours)	Llama 3	4	3	**86.54**	2.15
BayesVPGM (Ours)	Llama 3	2	3	85.49	1.81
BayesVPGM (Ours)	Llama 3	3	3	86.38	**1.05**
BayesVPGM (Ours)	Llama 3	4	3	**86.54**	<u>1.50</u>

Table 1: Accuracy (%) and ECE ($\times 10^2$) on ScienceQA for different methods, base models, and numbers of latent variables N. M is the number of sampled responses. The best and second-best results within each base model are **bolded** and <u>underlined</u>, respectively.

Model variant	Acc. ↑	ECE ↓
Chameleon+	85.17	8.65
w/ PGM (vPGM)	86.54	2.15
w/ PGM & nBI (BayesVPGM)	86.54	1.5

Table 2: Ablation on each model component tested on ScienceQA (accuracy in % and ECE in $\times 10^2$).

Table 2 isolates the contribution of each component. Adding the PGM alone markedly improves calibration (ECE 8.65 to 2.15) with an accuracy gain. The BayesVPGM, which combines PGM with nBI, retains the accuracy boost while achieving the lowest ECE, underscoring the importance of Bayesian refinement for reliable confidence estimation.

3.5 Results on A-OKVQA-clean

We present the results on A-OKVQA-clean data in Table 3. Our methods, vPGM and BayesVPGM, demonstrate performance on par with the baseline method Chameleon+.

Method	Acc.	ECE
Chameleon+	95.02	**2.75**
vPGM (Ours)	95.02	5.56
BayesVPGM (Ours)	95.02	<u>5.30</u>

Table 3: Performance on A-OKVQA-clean data (accuracy in % and ECE in $\times 10^2$).

3.6 Token-Level Computational Costs

We measured token-level costs on ScienceQA using Llama3-8B-Instruct; average instruction and output tokens per single query are shown in Table 4. All methods using multiple queries (e.g., $M = 3$ for Chameleon+ and vPGM variants) scale linearly in cost with M. While vanilla vPGM, BayesVPGM incur a higher per-query cost, the difference remains within the same order of magnitude. This is mainly due to their richer prompt structure and embedded probabilistic reasoning.

Method	N	Instruction	Output
CoT	-	627	83
Chameleon	-	1556	164
Chameleon+	-	1556	163
vPGM (Ours)	2	2394	458
BayesVPGM (Ours)	2	2394	458

Table 4: Token count for methods on ScienceQA, all methods are run using Llama 3.

3.7 Human Evaluation on ChatCoach

To complement the reference-based metric shown in Table 5 of the main paper, we conducted a human evaluation of all methods' final responses and vPGM's reasoning steps as shown in Table 5. Following the protocol of [3], two medically trained annotators rated each instance on a 1–4 scale for *constructiveness*, *clarity*, and *knowledgeability*. Because vPGM explicitly exposes its probabilistic reasoning, we also introduced an *interpretability* metric to assess how well those steps capture key task information. As Table 5 shows, vPGM achieves the highest scores in constructiveness, and knowledgeability, and remains competitive on clarity. We ensured annotation consistency by validating ratings, computing inter-annotator agreement with GPT-4, and flagged any instances where the two ratings diverged. Disagreements were resolved by joint review: annotators identified the source of confusion, clarified the guidelines, and re-annotated those items.

Table 5: Human evaluation on generation quality (126 instances, 10% of the test set) of methods for ChatCoach.

Metric	CoT	GCoT	vPGM (Ours)
Constructiveness	2.4	2.7	**2.8**
Clarity	2.2	**3.1**	2.8
Knowledgeability	2.4	2.4	**2.5**
Interpretability	-	-	**3.0**

3.8 Prompts and Examples

ScienceQA We provide a detailed example of inference using the vPGM, as shown in Table 6. Additionally, Table 7 demonstrates the prompt for a vPGM with 2 latent variables.

A-OKVQA Negative Control Here, we provide a detailed example of inference using the vPGM, as shown in Table 8. Additionally, Table 9 illustrates the inference prompt for vPGM with 2 latent variables.

ChatCoach We present the prompts used for each baseline approach: Instruction Prompting (see Table 10), Vanilla CoT (see Table 11), Zero-shot CoT (see Table 12), and GCoT (see Table 13). Table 14 shows the prompt of vPGM inference for ChatCoach.

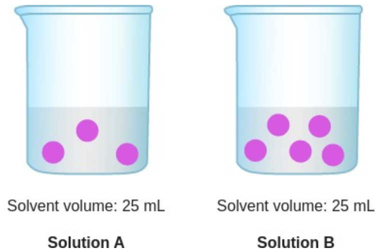
Question	Which solution has a higher concentration of pink particles? 
Answer options	(A) Neither; their concentrations are the same ✗ (B) Solution A ✗ (C) Solution B ✓
Image caption	A close up picture of a wii game controller.
Chameleon+	To determine which solution has a higher concentration of pink particles, we need some experimental data to compare their concentrations...Therefore, the answer is B. Probability: 0.852 *(Continue to obtain 2 more samples)* **Final Answer: (B) ✗** **Marginalized Probability Across 3 Samples:** $P(\mathbf{Y} = \text{A} \mid \mathbf{X}) = 0.852$
BayesVPGM	**Understanding Knowledge and Context (Variable Z_1):** Comparing the concentration of pink particles in two solutions...Given the lack of retrieved knowledge and Bing search response, the probability of Z_1 capturing the essential knowledge and context accurately is low. $P(Z_1 \mid \mathbf{X}) = 0.200$ **Analyzing Textual and Visual Information (Variable Z_2):** Detected Text: None provided. Image Caption: Mentions a wii game controller, which is not relevant to the question or the context...Given the lack of detected text and the irrelevant image caption, the probability of Z_2 accurately reflecting the meaning difference and assigning appropriate weightage is low. $P(Z_2 \mid Z_1, \mathbf{X}) = 0.200$ **Make a Decision with Confidence:** Option A (neither): Not supported. Without any information about the concentration of pink particles in the solutions, it cannot be concluded...Option B (Solution A): Not supported. Without any information about the concentration of pink particles...Option C (Solution B): Not supported...Given the lack of information about the concentration of pink particles in the solutions, the confidence in each option being correct is as follows. $P(\mathbf{Y} = \text{A} \mid Z_1, Z_2) = 0.333$ $P(\mathbf{Y} = \text{B} \mid Z_1, Z_2) = 0.333$ $P(\mathbf{Y} = \text{C} \mid Z_1, Z_2) = 0.333$ *(Obtain 2 more samples and perform Bayesian inference)* **Final Probabilities:** $P(\mathbf{Y} = \text{A} \mid Z_1, Z_2) = 0.073$ $P(\mathbf{Y} = \text{B} \mid Z_1, Z_2) = 0.416$ $P(\mathbf{Y} = \text{C} \mid Z_1, Z_2) = 0.510$ **Final Answer: (C) ✓**

Table 6: ScienceQA case study example. BayesVPGM captures discrepancies and produces a lower confidence comparing to Chameleon+.

Prompt for Inference of vPGM with 2 Latent Variables for ScienceQA
Perform step-by-step probabilistic reasoning given the following conditional distributions: 1. **Estimate** $P(Z_1 \mid \mathbf{X})$**:** Begin by analyzing the question and its associated context. Consider the relevance of external knowledge derived from knowledge retrieval outputs to the posed question. 2. **Estimate** $P(Z_2 \mid Z_1, \mathbf{X})$**:** Review the OCR results and image caption. Use the insights from Z_1 to evaluate the likelihood of discrepancies between the visual information and the textual context or question. 3. **Estimate** $P(\mathbf{Y} \mid Z_1, Z_2)$**:** Combine the assessments and reasoning derived from Z_1 and Z_2. For each answer option, assign a probability reflecting your confidence in its correctness.

Table 7: This is an example prompt for vPGM (with 2 latent variables) inference in the context of compositional reasoning tasks.

Question	What setting to the animals shown here prefer? 
Answer options	(A) Desert ✗ (B) Tropical ✗ (C) Savannah ✓ (D) Tundra ✗
Image caption	Two giraffes and other animals at a zoo.
Rationale generator	The man is wearing a sailor hat.
Chameleon+	Identify the preferred setting for the animals shown. The question asks about the preferred setting for the animals shown in the image...Therefore, the answer is B. Probability: 0.250 *(Continue to obtain 2 more samples)* **Final Answer: (B) ✗** **Marginalized Probability Across 3 Samples:** $P(\mathbf{Y} = \mathrm{B} \mid \mathbf{X}) = 0.425$
BayesVPGM	**Understanding Knowledge and Context (Variable Z_1):** Identifying the preferred habitat for the animals shown in the image...Given the complete lack of relevant information in the rationale, the probability that Z_1 captures the essential knowledge and context required for answering the question correctly is very low. $P(Z_1 \mid \mathbf{X}) = 0.050$ **Analyzing Textual and Visual Information (Variable Z_2):** The image caption mentions two giraffes and other animals at a zoo...Given the lack of specific details about the animals' preferred habitat in the image caption, the probability that Z_2 accurately reflects this information is low. $P(Z_2 \mid Z_1, \mathbf{X}) = 0.150$ **Make a Decision with Confidence:** Options A, B, D: These options are not directly supported by the information available, and there is no clear indication of the animals' preferred habitat. Option C (Savannah): While not explicitly mentioned, giraffes are commonly associated with Savannah habitats. $P(\mathbf{Y} = \mathrm{A} \mid Z_1, Z_2) = 0.100$ $P(\mathbf{Y} = \mathrm{B} \mid Z_1, Z_2) = 0.100$ $P(\mathbf{Y} = \mathrm{C} \mid Z_1, Z_2) = 0.650$ $P(\mathbf{Y} = \mathrm{D} \mid Z_1, Z_2) = 0.150$ *(Obtain 2 more samples and perform Bayesian inference)* **Final Probabilities:** $P(\mathbf{Y} = \mathrm{A} \mid Z_1, Z_2) = 0.179$ $P(\mathbf{Y} = \mathrm{B} \mid Z_1, Z_2) = 0.179$ $P(\mathbf{Y} = \mathrm{C} \mid Z_1, Z_2) = 0.440$ $P(\mathbf{Y} = \mathrm{D} \mid Z_1, Z_2) = 0.202$ **Final Answer: (C) ✓**

Table 8: A-OKVQA negative control case study example. BayesVPGM captures discrepancies between the rationale and the question, and produces the correct answer.

Prompt for Inference of vPGM with 2 Latent Variables for A-OKVQA
We have a question that requires a careful analysis to identify the correct answer. The decision-making process is structured into a series of steps, each focusing on specific aspects of the information provided. Let's approach this systematically: 1. **Estimate** $P(Z_1 \mid \mathbf{X})$**:** Start by analyzing the question and the provided context. What is the main topic, and what specific knowledge does it require? Consider the retrieved knowledge and the Bing search response. What essential information do these sources provide? Calculate the probability of Z_1 capturing the essential knowledge and context required for solving the question. 2. **Estimate** $P(Z_2 \mid Z_1, \mathbf{X})$**:** Examine the detected text in the image and the image caption. What are the key pieces of information each source provides? Are there any discrepancies between them? Estimate the probability of Z_2 accurately reflecting the meaning difference between detected text and image caption, and deciding the weight-age of each based on the discrepancy. 3. **Estimate** $P(\mathbf{Y} \mid Z_1, Z_2)$**:** Integrate the evaluations and reasoning from Z_1 to Z_2. For each answer option, provide a probability that represents your confidence in the option being correct. Ensure the probabilities sum up to 1. Proceed with the analysis, ensuring that each variable is logically derived from the provided information and the outcomes of the dependent variables.

Table 9: This is an example prompt for vPGM (with 2 latent variables) inference for A-OKVQA reasoning tasks.

Vanilla Instruction Prompting
Instruction: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. If there are discrepancies, guide the doctor. If not, provide positive feedback.

Table 10: Instruction prompting for ChatCoach.

Vanilla Chain-of-thought
Instruction: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. You should provide your response based on the following examples of input, thinking steps and output. **Example 1**: **Input**: {doctor's statement for Example 1} {medical context for Example 1} **Thinking steps**: {thinking steps for Example 1} **Output**: {coach's feedback for Example 1} **Example 2**: {example2} **Example 3**: {example3} **Input**: {doctor's statement} {medical context}

Table 11: Vanilla CoT for ChatCoach.

Zero-shot Chain-of-thought

Instruction: As a linguistic coach for a junior doctor, evaluate the doctor's statement: {doctor's statement} against the given medical context: {medical context}. If there are discrepancies, guide the doctor. If not, provide positive feedback.
Please think step by step.

Table 12: Zero-shot CoT for ChatCoach

Generalized Chain-of-thought (GCoT)

Instruction: As a linguistic coach for a junior doctor, your task is to evaluate the doctor's statement: {doctor's statement} against the provided medical context: {Medical Context}. Your evaluation should identify any discrepancies within the doctor's communication. Where discrepancies arise, guide the doctor towards more accurate medical terminology and understanding. If the statements align well with the medical context, provide positive reinforcement and additional advice if necessary.
Thinking steps:
Identify Key Medical Terms:
Extract medical terms from the doctor's statement, including diseases, symptoms, medications, and treatments.
Compare with Medical Context:
Check these terms against the medical context for accuracy in:

- Disease/symptom identification.
- Medication/treatment recommendation.

Feedback:

- *If Incorrect:* Point out the error and provide the correct term from the medical context. Use simple corrections like "Instead of [incorrect symptom], it should be [correct symptom]", "Instead of [incorrect medication name], it should be [correct medication name]" or "Instead of [incorrect disease name], it should be [correct disease name]".
- *If Correct:* Affirm with "Your diagnosis/treatment aligns well with the medical context. Good job."

Note: ¡correct symptom¿, ¡correct medication name¿ and ¡correct medication name¿ are extracted from medical context

Table 13: GCoT prompt for ChatCoach.

Prompt for vPGM inference for ChatCoach
Given the {doctor's statement} and the {medical context} provided: **Assess the Probability of Incorrect Terminology** ($P(Z_1)$): Analyze the medical terms used in the {doctor's statement}. Estimate the probability that any given medical term is used incorrectly based on the medical context. If medical term is irrelevant to medical context then it was considered incorrect. List the medical terms along with their corresponding numerical probability of being incorrect. **Identify Specific Errors** ($P(Z_2\|Z_1)$): For medical terms with a high probability of being incorrect, identify the specific term(s) that are used inappropriately. Provide a brief explanation for each identified error, referencing the {medical context}. **Determine Correction Requirement** ($P(Z_3\|Z_2, Z_1)$): Based on the errors identified, decide if a correction is needed for each term. For each term that requires correction, state the appropriate medical term extracted from medical context that should be used. For each step, provide your reasoning and the associated probabilities (give real numbers ranging from 0 to 1) , if applicable, to mimic the process of Bayesian inference. Conclude by generating the coach feedback (in Chinese) that assesses the doctor's statement against a provided medical context and guides the physician by pointing out the particular medical terminology errors and providing the corresponding corrections if discrepancies arise, if no mistakes occurred, then encouraging the doctor and provide further medical advice.

Table 14: Prompt of vPGM inference for ChatCoach

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[3] Hengguan Huang, Songtao Wang, Hongfu Liu, Hao Wang, and Ye Wang. Benchmarking large language models on communicative medical coaching: a novel system and dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[4] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.