Checklist

1. For all authors...
   - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   - (b) Did you describe the limitations of your work? [Yes] See 7
   - (c) Did you discuss any potential negative societal impacts of your work? [Yes] See 7
   - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   - (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See A.2
   - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See B
   - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported mean with standard errors
   - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   - (a) If your work uses existing assets, did you cite the creators? [N/A]
   - (b) Did you mention the license of the assets? [Yes] See A.3
   - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See A.2
   - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] As the data used in this study is publicly available online, explicit consent from individuals whose data is curated was not obtained. The data was accessed from publicly accessible sources, and no private or sensitive information was collected or utilized in this study.
   - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The dataset used in this study consists of question-answer pairs regarding scientific papers and does not explicitly contain any personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See A.5
   - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See A.5

A Dataset Documentation

A.1 Dataset Description

cPAPERS is a dataset of conversational question-answer pairs from reviews of academic papers grounded in these paper components and their associated references from scientific documents available on arXiv. For a detailed description and intended uses, please refer to 1.

A.2 Dataset Accessibility

- The dataset is publicly available at `https://huggingface.co/datasets/avalab/cPAPERS`
- URL to Croissant metadata record for viewing and downloading by the reviewers: `https://huggingface.co/datasets/avalab/cPAPERS/blob/main/croissant.json`
- Code repository for collecting the dataset: `https://github.com/jxu81/cPAPERS`
- Code repository for reproducing the benchmark results: `https://github.com/jxu81/cPAPERS`

A.3 Hosting, licensing, and maintenance plan

The dataset is licensed under Creative Commons (CC), and the code is licensed under the GNU General Public License (GPL). We plan to host and maintain this dataset on HuggingFace.

A.4 Dataset Examples

Example question-answer pairs are provided in Tables 9 10 11, .

	Example
Question	"What does the symbol ˜ mean in Equation 1?"
Answer	"The symbol ˜ in Equation 1 represents “follows this distribution”. It means that the probability distribution of the context C is defined as the distribution of the random variable C."
Question	"Can you provide more information about what is meant by ’generative process in biological multi-agent trajectories’ in L27 and L83?"
Answer	"The generative process refers to Eq. (2), which is a conceptual equation representing the generative process in animal behaviors."
Question	"How does the DeepMoD method differ from what is written in/after Eq 3?"
Answer	"We add noise only to u, with the percentage being of the standard deviation of the dataset. Adding noise to x and t has not been studied to our knowledge and falls out of the scope of this paper."
Question	"How to do the adaptive attack based on Eq.(16)? Maximizing the loss in Eq.(16)?"
Answer	"By Maximizing the loss in Eq (16) using an iterative method such as PGD on the end-to-end model we attempt to maximize the loss to cause misclassification while minimizing the regret to avoid detection."
Question	"How does the proposed method handle the imputed reward?"
Answer	"The proposed method uses the imputed reward in the second part of Equation 1, which corresponds to the empirical risk of the combined dataset."

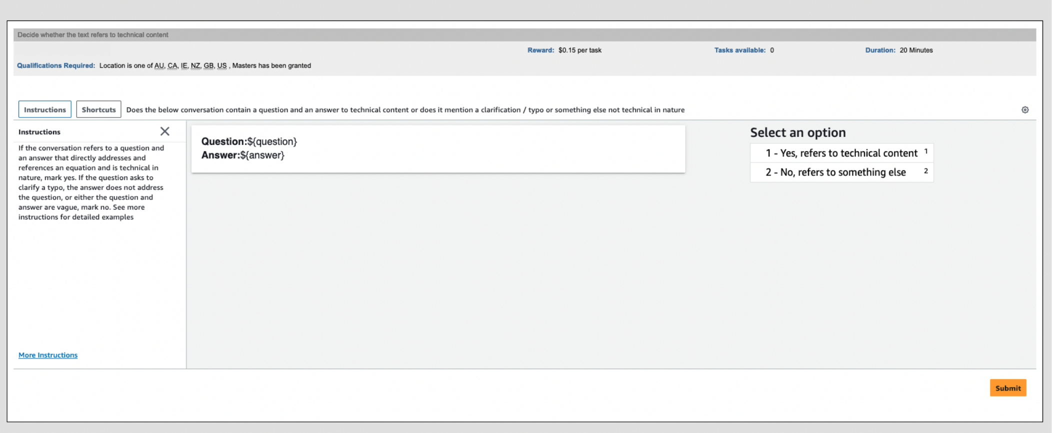
Table 9: Example QA Pairs in the cPAPERS-EQNS dataset

A.5 Crowdworker Instructions

A significant portion of academic reviews and rebuttals pertain to clarification questions and fixing typos. While the LLM processing removes most of these spurious questions and answers, to further ensure the quality of the dataset we employ crowdworkers from Amazon Mechanical Turk to ascertain whether a question-answer pair about an equation, table, or figure is technical in nature or asks to fix

	Example
Question	"What is the purpose of Table 2 in the paper?"
Answer	"Table 2 is used to provide a comparison of the computational complexity of the proposed approach with state-of-the-art methods."
Question	"Optimal number of clusters affected by the number of classes or similarity between classes?"
Answer	"The authors have addressed this concern by including a new experiment in Table 4 of the revised paper's appendix. The result shows that the optimal number of clusters is less affected by the number of classes but more affected by the similarity between classes."
Question	"Can you clarify the values represented in Table 1?"
Answer	"The values in Table 1 represent the number of evasions, which shows the attack strength. Therefore, the higher the value, the stronger the attack."
Question	"The experiments in table 1 do not seem to favor the proposed method much; softmax is better or similar. Can the authors explain why this might be the case?"
Answer	"The proposed method reduces to empirical risk minimization with a proper loss, and the seemingly trivial solution ϕ=Θ is often not optimal. This could explain why the proposed method might not perform as well as other methods in certain experiments. However, the authors hope that addressing concerns about the method's theoretical properties would be beneficial."
Question	"Does the first row of Table 2 correspond to the offline method?"
Answer	"Yes, the first row of Table 2 corresponds to the offline method."

Table 10: Example QA Pairs in the cPAPERS-TBLS dataset



Decide whether the text refers to technical content

Reward: $0.15 per task **Tasks available:** 0 **Duration:** 20 Minutes

Qualifications Required: Location is one of AU, CA, IE, NZ, GB, US , Masters has been granted

Does the below conversation contain a question and an answer to technical content or does it mention a clarification / typo or something else not technical in nature

Instructions

If the conversation refers to a question and an answer that directly addresses and references an equation and is technical in nature, mark yes. If the question asks to clarify a typo, the answer does not address the question, or either the question and answer are vague, mark no. See more instructions for detailed examples

Question:${question}
Answer:${answer}

Select an option

1 - Yes, refers to technical content 1
2 - No, refers to something else 2

Figure 2: Screenshot of crowdworker interface on Amazon Mechanical Turk

a typo. We recruit crowdworkers with the masters qualification from predominantly English-speaking countries: Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States of America. Crowdworkers are compensated USD 0.15 per Human Intelligence Task, with each task taking an average of one minute to complete. The crowdworker instructions and interface are provided in Figure 2. The total cost of cleaning up the dataset including fees paid to Amazon Mechanical Turk is USD 3215. Since the dataset is collected from OpenReview, we did not identify any risks to crowdworkers.

	Example
Question	"Why is there a gap between the proposed approach and the median approach in Fig. 5?"
Answer	"The gap is due to lower accuracy of the approximate median calculated by the bucketing scheme compared to a median method, albeit with faster epoch completion times."
Question	"What do experiments ensure us that the dependency is linear?"
Answer	"The linear dependency can be observed empirically in Figure 1. The paper will provide further experimental results in the updated version."
Question	"How do the different methods in Figure 5 have similar test errors, but the generalization bounds are so different?"
Answer	"The different methods in Figure 5 have similar test errors because they are all trained on the same dataset and have similar performance. However, the generalization bounds are different because they are computed using different methods. The proposed method uses a tighter bound that takes into account the structure of the ensemble, while the baselines use a looser bound that is based on an upper bound of the error rate."
Question	"What is the semantic meaning of “average episodic coverage” in Figure 5?"
Answer	"The semantic meaning of “average episodic coverage” in Figure 5 refers to the number of unique avatar positions. The authors have added a DIAYN baseline and a random agent baseline to provide context for how other methods fare."
Question	"In Figure 3, does the number of epochs mean the same thing for BAIL+ and MBAIL?"
Answer	"Yes, the number of epochs in Figure 3 means the same thing for BAIL+ and MBAIL. For MBAIL, it refers to the constant E in algorithm 2."

Table 11: Example QA Pairs in the cPAPERS-FIGS dataset

B Language Modeling Details

We randomly split the dataset into training (60%), dev set (20%), and test set (20%). Hyperparameters for parameter-efficient fine-tuning can be found in 12. Other model training details are listed in 13.

Parameter	Value
Rank	64
α	16
Dropout	0.1

Table 12: Hyperparameters for Parameter-efficient fine-tuning using QLoRA

Parameter	Value
Learning Rate	2e-4
Batch size	4
Warmup Schedule	Constant
Warmup Ratio	0.03
Epochs	5
Optimizer	`paged_adamw_32bit`[8]
Compute	8 Nvidia A40 GPUs

Table 13: Additional hyperparameters for fine-tuning experiments

B.1 Additional Model Performance

We conducted additional experiments to benchmark the baseline performance of state-of-the-art pre-trained LLMs in answering questions in the cPAPERS dataset without additional fine-tuning. In Tables 14, 15, and 16, we report the zero-shot performance of LLAMA-2-7B, LLAMA-2-70B, LLAMA-3-8B, and LLAMA-3-70B on the cPAPERS dataset.

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**	**BERTScore**
LLAMA-2-7B	0.189	0.060	0.136	0.232	0.823
LLAMA-2-70B	0.194	0.065	0.144	0.240	0.825
LLAMA-3-8B	0.139	0.047	0.107	0.161	0.768
LLAMA-3-70B	0.266	0.104	0.203	0.243	0.844

Table 14: Comparison of zero-shot performance across different models on the cPAPERS-EQNS test set

B.2 Impact of Temperatures on Zero-shot Language Modeling Results

We conducted additional experiments to evaluate the influence of temperature on the baseline performance. Temperature parameters were sampled from 0.0 to 1.0 at intervals of 0.1, and each experiment was repeated five times with five randomly generated seeds. Average scores and standard errors across metrics were computed, and the results are presented in tables 17, 18, and 19.

B.3 Impact of Temperatures on Fine-tuning Language Modeling Results

We conducted supplementary experiments to evaluate the influence of temperature on the fine-tuned model. Temperature parameters were sampled from 0.0 to 1.0 at intervals of 0.1, and each experiment was repeated five times with five randomly generated seeds. Average scores and standard errors across metrics were computed, and the results are presented in tables 20, 21, and 20.

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
LLAMA-2-7B	0.190	0.054	0.131	0.227	0.830
LLAMA-2-70B	0.192	0.058	0.136	0.232	0.832
LLAMA-3-8B	0.132	0.039	0.096	0.147	0.763
LLAMA-3-70B	0.256	0.086	0.187	0.217	0.850

Table 15: Comparison of zero-shot performance across different models on the cPAPERS-TBLS test set

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
LLAMA-2-7B	0.187	0.061	0.136	0.237	0.831
LLAMA-2-70B	0.185	0.065	0.137	0.238	0.833
LLAMA-3-8B	0.126	0.045	0.100	0.174	0.784
LLAMA-3-70B	0.282	0.119	0.218	0.256	0.853

Table 16: Comparison of zero-shot performance across different models on the cPAPERS-FIGS test set

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.194 ± 0.000	0.065 ± 0.000	0.144 ± 0.000	0.240 ± 0.000	0.825 ± 0.000
0.1	0.193 ± 0.001	0.064 ± 0.000	0.143 ± 0.000	0.238 ± 0.000	0.825 ± 0.000
0.3	0.194 ± 0.001	0.065 ± 0.001	0.143 ± 0.001	0.240 ± 0.000	0.825 ± 0.000
0.5	0.194 ± 0.001	0.065 ± 0.000	0.142 ± 0.000	0.240 ± 0.000	0.825 ± 0.000
0.7	0.193 ± 0.001	0.065 ± 0.001	0.142 ± 0.001	0.239 ± 0.000	0.825 ± 0.000
0.9	0.193 ± 0.001	0.064 ± 0.000	0.141 ± 0.001	0.240 ± 0.001	0.825 ± 0.000

Table 17: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-EQNS test set.

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.192 ± 0.000	0.058 ± 0.000	0.136 ± 0.000	0.232 ± 0.000	0.832 ± 0.000
0.1	0.192 ± 0.001	0.058 ± 0.000	0.136 ± 0.001	0.231 ± 0.000	0.832 ± 0.000
0.3	0.191 ± 0.001	0.058 ± 0.000	0.136 ± 0.000	0.232 ± 0.000	0.832 ± 0.000
0.5	0.192 ± 0.000	0.059 ± 0.000	0.137 ± 0.000	0.233 ± 0.001	0.832 ± 0.000
0.7	0.192 ± 0.001	0.058 ± 0.001	0.137 ± 0.001	0.233 ± 0.000	0.832 ± 0.000
0.9	0.191 ± 0.001	0.057 ± 0.001	0.136 ± 0.001	0.232 ± 0.000	0.832 ± 0.001

Table 18: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-TBLS test set.

Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
0.0	0.185 ± 0.000	0.065 ± 0.000	0.137 ± 0.000	0.238 ± 0.000	0.833 ± 0.000
0.1	0.188 ± 0.001	0.066 ± 0.000	0.140 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.3	0.189 ± 0.001	0.067 ± 0.001	0.141 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.5	0.191 ± 0.001	0.068 ± 0.001	0.143 ± 0.001	0.242 ± 0.001	0.835 ± 0.000
0.7	0.190 ± 0.001	0.067 ± 0.000	0.142 ± 0.000	0.241 ± 0.001	0.834 ± 0.000
0.9	0.190 ± 0.001	0.066 ± 0.001	0.141 ± 0.000	0.240 ± 0.000	0.834 ± 0.000

Table 19: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperature on the cPAPERS-FIGS test set.

Modality	**Temp**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**	**BERTScore**
Question (Q)	0.0	0.309 ± 0.000	0.138 ± 0.000	0.248 ± 0.000	0.215 ± 0.000	0.860 ± 0.000
	0.1	0.309 ± 0.001	0.139 ± 0.001	0.249 ± 0.000	0.216 ± 0.000	0.860 ± 0.001
	0.3	0.306 ± 0.001	0.135 ± 0.000	0.245 ± 0.001	0.213 ± 0.000	0.858 ± 0.002
	0.5	0.301 ± 0.001	0.132 ± 0.001	0.242 ± 0.000	0.212 ± 0.001	0.860 ± 0.001
	0.7	0.297 ± 0.001	0.125 ± 0.001	0.236 ± 0.001	0.210 ± 0.001	0.858 ± 0.001
	0.9	0.285 ± 0.002	0.113 ± 0.001	0.223 ± 0.001	0.202 ± 0.002	0.856 ± 0.001
Q + Equation	0.0	0.317 ± 0.000	0.134 ± 0.000	0.251 ± 0.000	0.223 ± 0.000	0.861 ± 0.000
	0.1	0.313 ± 0.001	0.133 ± 0.001	0.250 ± 0.001	0.220 ± 0.001	0.860 ± 0.001
	0.3	0.310 ± 0.001	0.130 ± 0.001	0.245 ± 0.001	0.221 ± 0.001	0.859 ± 0.002
	0.5	0.305 ± 0.002	0.124 ± 0.001	0.239 ± 0.002	0.217 ± 0.002	0.857 ± 0.001
	0.7	0.297 ± 0.002	0.116 ± 0.001	0.231 ± 0.001	0.213 ± 0.002	0.857 ± 0.001
	0.9	0.281 ± 0.002	0.102 ± 0.002	0.215 ± 0.001	0.206 ± 0.002	0.850 ± 0.001
Q + Context	0.0	0.297 ± 0.000	0.126 ± 0.000	0.235 ± 0.000	0.221 ± 0.000	0.817 ± 0.000
	0.1	0.297 ± 0.001	0.126 ± 0.001	0.235 ± 0.000	0.220 ± 0.001	0.820 ± 0.001
	0.3	0.297 ± 0.002	0.123 ± 0.001	0.234 ± 0.001	0.222 ± 0.001	0.824 ± 0.001
	0.5	0.297 ± 0.002	0.119 ± 0.001	0.232 ± 0.002	0.221 ± 0.001	0.832 ± 0.002
	0.7	0.287 ± 0.001	0.108 ± 0.002	0.219 ± 0.003	0.218 ± 0.001	0.832 ± 0.002
	0.9	0.270 ± 0.002	0.089 ± 0.003	0.200 ± 0.004	0.207 ± 0.002	0.826 ± 0.001
Q + References	0.0	0.283 ± 0.000	0.122 ± 0.000	0.224 ± 0.000	0.217 ± 0.000	0.777 ± 0.000
	0.1	0.286 ± 0.001	0.123 ± 0.001	0.226 ± 0.001	0.222 ± 0.001	0.785 ± 0.002
	0.3	0.283 ± 0.001	0.118 ± 0.001	0.222 ± 0.001	0.221 ± 0.001	0.786 ± 0.003
	0.5	0.281 ± 0.002	0.115 ± 0.001	0.219 ± 0.001	0.221 ± 0.002	0.785 ± 0.003
	0.7	0.271 ± 0.001	0.106 ± 0.001	0.208 ± 0.002	0.216 ± 0.002	0.781 ± 0.002
	0.9	0.258 ± 0.001	0.093 ± 0.001	0.193 ± 0.002	0.209 ± 0.002	0.770 ± 0.002

Table 20: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperatures on the cPAPERS-EQNS test set.

Modality	Temp	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
Question (Q)	0.0	0.315 ± 0.000	0.121 ± 0.000	0.235 ± 0.000	0.218 ± 0.000	0.869 ± 0.000
	0.1	0.316 ± 0.001	0.122 ± 0.001	0.234 ± 0.000	0.220 ± 0.001	0.869 ± 0.001
	0.3	0.313 ± 0.001	0.120 ± 0.001	0.232 ± 0.001	0.220 ± 0.001	0.869 ± 0.001
	0.5	0.309 ± 0.001	0.116 ± 0.001	0.229 ± 0.001	0.216 ± 0.001	0.868 ± 0.000
	0.7	0.302 ± 0.002	0.108 ± 0.002	0.221 ± 0.001	0.213 ± 0.001	0.868 ± 0.001
	0.9	0.288 ± 0.003	0.095 ± 0.002	0.208 ± 0.002	0.205 ± 0.003	0.865 ± 0.000
Q + Table	0.0	0.293 ± 0.000	0.107 ± 0.000	0.218 ± 0.000	0.212 ± 0.000	0.820 ± 0.000
	0.1	0.288 ± 0.002	0.105 ± 0.001	0.214 ± 0.001	0.211 ± 0.001	0.815 ± 0.003
	0.3	0.289 ± 0.002	0.103 ± 0.001	0.214 ± 0.001	0.215 ± 0.002	0.837 ± 0.003
	0.5	0.286 ± 0.001	0.096 ± 0.001	0.207 ± 0.001	0.216 ± 0.001	0.844 ± 0.003
	0.7	0.277 ± 0.002	0.085 ± 0.001	0.196 ± 0.001	0.212 ± 0.004	0.844 ± 0.003
	0.9	0.249 ± 0.002	0.064 ± 0.001	0.170 ± 0.001	0.200 ± 0.004	0.833 ± 0.004
Q + Context	0.0	0.294 ± 0.000	0.106 ± 0.000	0.216 ± 0.000	0.225 ± 0.000	0.838 ± 0.000
	0.1	0.301 ± 0.001	0.108 ± 0.000	0.219 ± 0.000	0.227 ± 0.000	0.846 ± 0.002
	0.3	0.301 ± 0.001	0.108 ± 0.001	0.218 ± 0.001	0.227 ± 0.001	0.846 ± 0.001
	0.5	0.297 ± 0.002	0.104 ± 0.001	0.214 ± 0.002	0.226 ± 0.002	0.848 ± 0.002
	0.7	0.287 ± 0.002	0.092 ± 0.001	0.203 ± 0.002	0.218 ± 0.003	0.845 ± 0.002
	0.9	0.272 ± 0.002	0.079 ± 0.001	0.187 ± 0.002	0.213 ± 0.003	0.842 ± 0.001
Q + Reference	0.0	0.292 ± 0.000	0.106 ± 0.000	0.214 ± 0.000	0.218 ± 0.000	0.816 ± 0.000
	0.1	0.297 ± 0.001	0.108 ± 0.001	0.217 ± 0.001	0.222 ± 0.001	0.830 ± 0.002
	0.3	0.298 ± 0.002	0.105 ± 0.001	0.217 ± 0.001	0.224 ± 0.001	0.842 ± 0.002
	0.5	0.294 ± 0.002	0.101 ± 0.001	0.212 ± 0.002	0.221 ± 0.001	0.843 ± 0.003
	0.7	0.288 ± 0.001	0.092 ± 0.001	0.205 ± 0.001	0.219 ± 0.001	0.847 ± 0.001
	0.9	0.272 ± 0.002	0.081 ± 0.001	0.190 ± 0.001	0.212 ± 0.002	0.837 ± 0.002

Table 21: Zero-shot performance (mean and standard errors over 5 seeds) of Llama-2-70B across different temperatures on the cPAPERS-TBLS test set.

Modality	**Temp**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**	**BERTScore**
Question (Q)	0.0	0.329 ± 0.000	0.155 ± 0.000	0.269 ± 0.000	0.250 ± 0.000	0.859 ± 0.000
	0.1	0.333 ± 0.001	0.157 ± 0.001	0.272 ± 0.001	0.252 ± 0.001	0.859 ± 0.002
	0.3	0.329 ± 0.002	0.153 ± 0.001	0.267 ± 0.001	0.249 ± 0.001	0.860 ± 0.003
	0.5	0.324 ± 0.002	0.145 ± 0.002	0.259 ± 0.002	0.244 ± 0.002	0.861 ± 0.002
	0.7	0.315 ± 0.002	0.134 ± 0.002	0.248 ± 0.002	0.242 ± 0.002	0.862 ± 0.001
	0.9	0.291 ± 0.001	0.113 ± 0.002	0.223 ± 0.001	0.232 ± 0.002	0.857 ± 0.003
Question + Figure	0.0	0.322 ± 0.000	0.147 ± 0.000	0.260 ± 0.000	0.246 ± 0.000	0.868 ± 0.000
	0.1	0.325 ± 0.001	0.149 ± 0.001	0.262 ± 0.001	0.246 ± 0.001	0.865 ± 0.001
	0.3	0.322 ± 0.002	0.146 ± 0.001	0.258 ± 0.002	0.245 ± 0.002	0.865 ± 0.001
	0.5	0.318 ± 0.001	0.139 ± 0.001	0.251 ± 0.001	0.242 ± 0.001	0.865 ± 0.001
	0.7	0.306 ± 0.001	0.125 ± 0.001	0.239 ± 0.001	0.238 ± 0.002	0.863 ± 0.001
	0.9	0.281 ± 0.001	0.102 ± 0.002	0.212 ± 0.001	0.226 ± 0.001	0.852 ± 0.003
Context	0.0	0.321 ± 0.000	0.140 ± 0.000	0.256 ± 0.000	0.250 ± 0.000	0.858 ± 0.000
	0.1	0.317 ± 0.002	0.138 ± 0.001	0.254 ± 0.001	0.248 ± 0.001	0.853 ± 0.002
	0.3	0.323 ± 0.002	0.138 ± 0.001	0.255 ± 0.001	0.252 ± 0.001	0.861 ± 0.001
	0.5	0.318 ± 0.001	0.132 ± 0.002	0.248 ± 0.002	0.250 ± 0.002	0.863 ± 0.001
	0.7	0.307 ± 0.001	0.122 ± 0.002	0.236 ± 0.002	0.242 ± 0.002	0.861 ± 0.001
	0.9	0.278 ± 0.001	0.097 ± 0.002	0.209 ± 0.002	0.225 ± 0.002	0.851 ± 0.001
Reference	0.0	0.311 ± 0.000	0.131 ± 0.000	0.244 ± 0.000	0.247 ± 0.000	0.843 ± 0.000
	0.1	0.314 ± 0.002	0.134 ± 0.001	0.248 ± 0.001	0.250 ± 0.001	0.848 ± 0.003
	0.3	0.310 ± 0.002	0.130 ± 0.001	0.244 ± 0.001	0.249 ± 0.002	0.845 ± 0.002
	0.5	0.308 ± 0.002	0.125 ± 0.001	0.240 ± 0.001	0.249 ± 0.002	0.857 ± 0.003
	0.7	0.296 ± 0.001	0.113 ± 0.001	0.226 ± 0.001	0.241 ± 0.001	0.854 ± 0.002
	0.9	0.269 ± 0.001	0.090 ± 0.001	0.198 ± 0.002	0.226 ± 0.002	0.843 ± 0.001

Table 22: Zero-shot performance (mean and standard errors over 5 seeds) of LLAMA-2-70B across different temperatures on the cPAPERS-FIGS test set.