How (not) to ensemble LVLMs for VQA

Lisa Alazraki[*]
Imperial College London

Lluis Castrejon
Google Research

Mostafa Dehghani
Google DeepMind

Fantine Huot
Google DeepMind

Jasper Uijlings
Google Research

Thomas Mensink
Google Research

Abstract

This paper studies ensembling in the era of Large Vision-Language Models (LVLMs). Ensembling is a classical method to combine different models to get increased performance. In the recent work on Encyclopedic-VQA [18] the authors examine a wide variety of models to solve their task: from vanilla LVLMs, to models including the caption as extra context, to models augmented with Lens-based retrieval of Wikipedia pages. Intuitively these models are highly complementary, which should make them ideal for ensembling. Indeed, an oracle experiment (Fig. 1) shows potential gains from 48.8% accuracy (the best single model) all the way up to 67% (best possible ensemble). So it is a trivial exercise to create an ensemble with substantial real gains. Or is it?

1 Introduction

Large Vision-Language models (LVLMs) have achieved impressive results on Visual Question Answering (VQA) [2, 6, 13, 28]. Ensembling multiple LVLMs has the potential to increase performance even further. In this work, we focus on Encyclopedic-VQA – a recently-introduced VQA task [18] asking questions about detailed properties of finegrained categories (Fig. 2). In [18] they use nine LVLMs: PaLI [6], PaLM [7] and GPT-3 [4], each deployed with three different augmentation strategies: (1) PromptCap [14], (2) Wikipedia sections retrieved via Google Lens [1] and (3) without any augmentation ('vanilla'). Intuitively, these approaches are quite different and complementary, which makes them ideal for ensembling from a theoretical perspective [8, 27]. Therefore in this paper we set out to create a strong ensemble out of these models.

To show the potential of ensembling, we carry out an oracle experiment on Encyclopedic-VQA (Fig. 1), on the single-hop, single-answer questions of the dataset [18]. In particular, we select for each VQA example the best answer out of those given by the nine LVLMs. Whereas the best single model achieves 48.8% accuracy, the best possible ensemble achieves 67.0% accuracy. This demonstrates we have a large potential gain of 18.2%!

Hence we explore in this paper a variety of ensembling methods: classical ensembling techniques such as majority voting and using model confidence (Sec. 2), prompting the LVLMs to do self-reflection (Sec. 3), and finally using an external evaluator model to judge which answer is correct (Sec. 4). In our exploration we aim to find which ensembling techniques work, which do not, and why.

Our contributions are the following: (1) we identify the large potential gain of ensembling multiple LVLMs on Encyclopedic-VQA; (2) we explore a variety of ensembling methods for LVLMs: classical methods, self-reflection, and external evaluation by another LVLM; (3) we increase performance on Encyclopedic-VQA by 4.6% through a model cascade using external evaluation. Our analysis

[*]Work done during an internship at Google.
Contact: `lisa.alazraki20@imperial.ac.uk`, or `{jrru,mensink}@google.com`,

4th *I Can't Believe It's Not Better* Workshop (co-located with NeurIPS 2023)

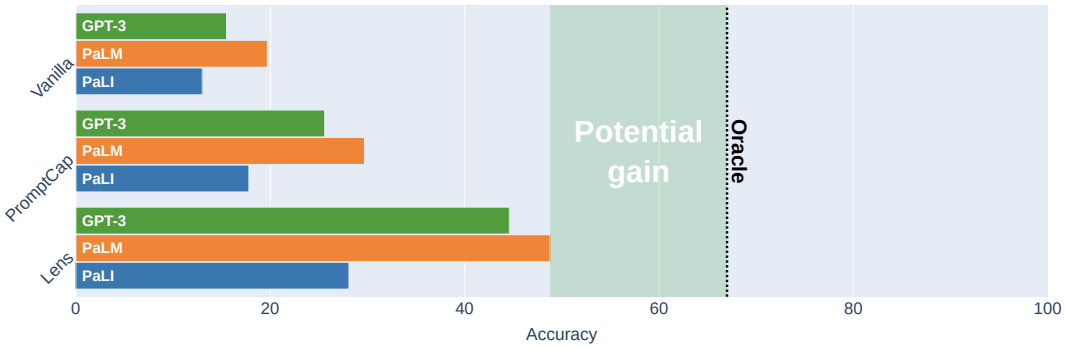

Figure 1: Encyclopedic-VQA baselines and oracle ensemble of nine LVLMs. All results are on the single-hop single-answer questions following the main experiments in [18].

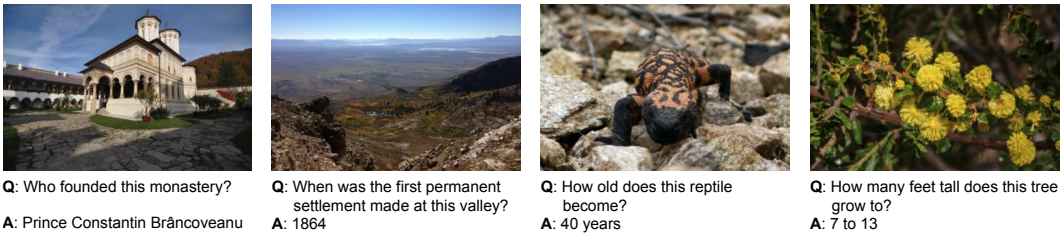

Figure 2: Examples from the Encyclopedic-VQA task.

shows that the majority of this boost is attributed to the utilization of a larger model for evaluation and a singular critical observation; (4) most of the potential gain remains untapped and most of our ensemble strategies were not as successful as envisioned. This leads us to conclude that effectively ensembling LVLMs is challenging.

2 Classical Ensembling Methods

We begin our investigation into ensembling from established methods that have been widely used before: majority voting [11, 21, 15] and using the model's own confidence [17, 10, 22]. Before doing so, we detail our oracle experiment.

Oracle Ensembling We run inference on Encyclopedic-VQA using the nine LVLMs in [18]. The oracle selector picks one of the LVLMs yielding the correct answer, or a random one if none of the LVLMs has produced the correct answer. We evaluate the oracle selector on the test set and obtain 67.0% BEM accuracy [5]. The oracle thus shows a large potential improvement of 18.2% over the accuracy achieved by PaLM with Lens (48.8%), the best-performing single LVLM in [18].

2.1 Majority Voting

In majority voting, we use the most voted answer among the outputs of the nine LVLMs. We adopt BEM-style soft-matching (*i.e.* using BERT matching [5] to determine whether two answers are the same). We evaluate majority voting on the Encyclopedic-VQA test set and obtain 45.3%, much lower than the best single model (48.8%). Hence we conclude that majority voting does not improve the VQA performance for this task.

2.2 Model Confidence

In this section we focus on using the sequence probabilities of the LVLMs as a signal for ensembling. Sequence probabilities have been used to estimate confidence in a QA answer in [24]. Similarly to [24], we normalize the probabilities for sequence length, *i.e.* $\bar{p} = p^{\frac{1}{N}}$, with $p = \prod_{i=0}^{N} p(t_i|t_{<i})$ for a sequence of N tokens.

Ensemble methods. For the following experiments we use three LVLMs: PaLI, PaLM, and GPT-3 all with the Lens retrieval setup, for a subset of 1,000 Encyclopedic-VQA examples. We attempt three different strategies to ensemble these models according to their (normalized) probabilities:

Table 1: Results on the Lens retrieval setup.

	Acc. (BEM)
PaLM w/ Lens (baseline)	49.4%
Choosing max prob. answer	49.2%
Weighted voting	49.2%
Logistic regression	50.6%
Oracle ensemble	57.5%

(a) Classical ensembling results.

	PaLI	PaLM	GPT-3
BEM (↑)	34.6%	49.4%	47.1%
ECE (↓)	0.121	0.175	0.272
Brier (↓)	0.182	0.272	0.278
ECE-re (↓)	0.039	0.131	0.127
Brier-re (↓)	0.164	0.256	0.221

(b) Calibration metrics.

1. **Max probability.** Choose per example the LVLM with the highest probability.
2. **Weighted voting.** If the same answer is given by multiple LVLMs use the average probability across the LVLMs, otherwise the probability of the LVLM producing the answer.
3. **Classification.** Train a classifier to learn the weights for weighted voting, using the probabilities of all three models as inputs.

The resulting BEM scores (Tab. 1a), show that all strategies perform similarly to the baseline. The logistic regression classifier slightly outperforms it (by 1.2%), yet its BEM score is far below that achieved by the oracle ensemble of those three models. We also try adding to the classifier feature transformations (such as Z-score normalisation, sqrt, or power), feature combinations (*e.g.* $p1 \times p2$), and hidden layers in an MLP to add more expressivity, all without much results. We hence conclude that LVLM sequence probabilities are not reliable enough to be used for ensembling. In the next paragraph we discuss if this could be due to miscalibration.

Calibration. In order to understand if the sequence probabilities of the different LVLMs are comparable, we investigate here whether these are calibrated w.r.t. VQA accuracy. In other words, we examine whether these generative probabilities are also representative of the likelihood of an answer being correct. Perfect calibration would mean that if the probability is 80%, we find (empirically) that the answer is also correct 80% of the times.

To measure calibration we use the ECE [20] and the Brier score [3] as metrics, which we evaluate over the same subset of examples and LVLMs as above. The results are in Tab. 1b. We observe (1) that the accuracy of the chosen examples is slightly above the average; (2) that the Brier score and the ECE metric are correlated, *i.e.* that the method with the lowest Brier score also has the lowest ECE, (3) that the PaLI LVLM is best calibrated, but does not yield the highest BEM score.

We also optimize the ECE/Brier scores by temperature scaling, *i.e.* we derive p from the sequence log-likelihood l as $p = \exp l/t$, where the temperature t is individually tuned for each LVLM (see Appendix C). We find that this method of calibration significantly improves the ECE and Brier scores as evidenced in Tab. 1b. However, ensembling via model selection using the maximum of the calibrated probabilities performs identically to the non-calibrated probabilities. We hence conclude that while temperature scaling leads to better calibrated LVLMs, their sequence probabilities remain a weak signal for ensembling.

Limitations. In this section we use sequence probabilities as a measure of the model's confidence in VQA answers, following [24]. The relatively low ECE and Brier scores justify this approach. However, recent work [16] criticizes this method, as different yet semantically equivalent answers would yield different probabilities despite being equally correct. Future work could investigate the use of semantic likelihood [16] in place of sequence probability.

Another limitation is that for typical weighted voting methods one averages the models over each class. For LVLMs something similar could be achieved by first collecting all answers, and then have each LVLM output the probability of each of these answers. Unfortunately, due to the limitations of the APIs this was not possible to do for all models and we could not explore this avenue.

3 Self-Reflection

LVLMs open up a new and interesting way of obtaining a confidence measure by having them self-reflect on their own prediction through prompting [25, 26]. We explore this here.

In preliminary experiments we tried prompting the LVLMs to output a numerical value for their confidence, as previously explored in [26]. These experiments showed that these models are not good

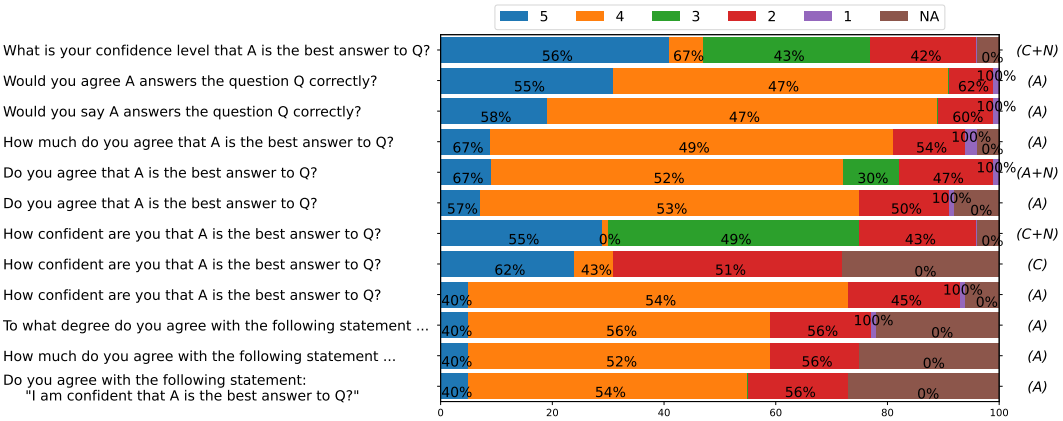

Figure 3: Variations of the same prompt to elicit a Likert-Scale confidence prediction, evaluated using PaLM on 100 examples. The width of the bars shows the proportion of examples falling in each scale. The percentage represents the BEM accuracy of the questions in each Likert-bucket, which ideally should be high when the model is very confident abouts its answer (scale 5 in blue) and low when the model is very unconfident about its answer (scale 1 in purple). 'C' means our Likert scale is phrased in confidence: 'very confident', 'confident', 'neither confident nor unconfident', 'unconfident', 'very unconfident'. 'A' indicates the same scale but in terms of 'agree'. For 'N' we rephrased the middle option (in green) as 'neutral'.

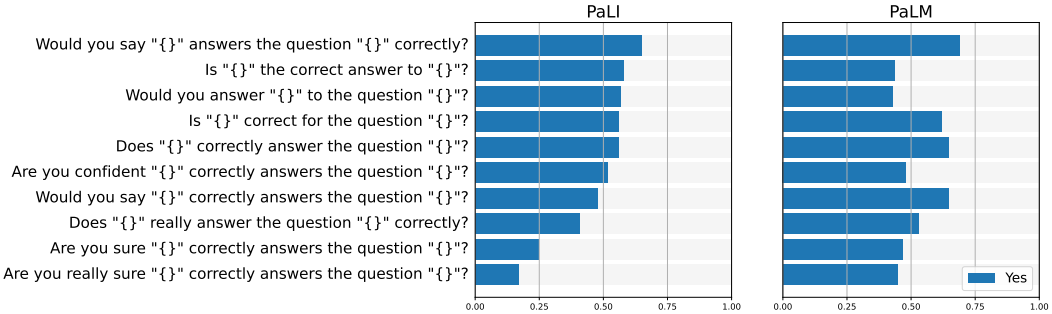

Figure 4: Evaluation of different prompts to get a yes/no confidence reflection. Ten different prompts are evaluated for PaLI and PaLM on the same subset of 100 examples. We observe a large variation in the distribution of the yes/no responses between the different prompts, and across the two models.

at providing a numerical value for complex concepts such as confidence or correctness. Hence we focus on language variations of confidence estimations, similarly to [25], using yes/no questions and the 5-point Likert scale.

Likert scale. Prompts including a 5-point Likert scale are too long for PaLI, so we perform these experiments using PaLM only. Fig. 3 shows several variations of the same prompt to elicit a confidence score. In this figure, the size of the bars corresponds to the percentage of answers which fall in a confidence bucket: 5 (blue) for very confident, and 1 (purple) for very unconfident. We also plot the corresponding BEM accuracy scores averaged over the answers in each bucket as a percentage on top of it. Now examining the first prompt (top), we see that the BEM accuracy is 56% for answers where the LVLM is 'very confident'. For answers where it is just 'confident', the BEM accuracy increases to 67% whereas we would expect a decrease if the self-reflection was good. Results are even more dramatic for the second row: BEM accuracy is highest (62%) for answers which the model is 'unconfident' about (!). Hence for our problem self-reflection does not seem very accurate.

Another observations is that the distribution of answers varies quite a lot in Fig. 3, as can be seen by the different proportions (widths) of colors within each bar. This is problematic, since it suggests that the task of self-reflection is less important than how the the self-reflection prompt is phrased.

Finally, we note that depending on how the questions are phrased, quite a few answers fall outside the expected scale (NA in brown, Fig. 3); the *neither agree nor disagree* was hard for the model to produce. However, when we made this option simpler by phrasing it as *neutral* it also affected how many times other Likert scale points were selected (compare row 7 (C+N) with row 8 (C)).

Binary responses. To obtain a binary confidence score, we use a prompt asking the LVLM about its confidence and requiring a yes/no response, *e.g.* `Are you confident that "A" correctly answers "Q"? Output yes or no`. In preliminary results we found that the obtained binary scores do not correlate highly with the correctness of an answer (data not shown). Hence binary self-reflection did not work for us either. To understand why, we again test several paraphrases of the same prompt and plot the distribution of answers in Fig. 4. Again, we see that the exact phrasing of the prompt seems to be the dominant factor in these results. For example, using the top question PaLI is confident of its answer in 65% of the cases, whereas using the bottom question it is confident only in 17% of the cases. For PaLM, results vary between 40%-69%, which is comparatively less but still a worrying amount of variation.

Overall, we conclude that the phrasing of the prompt dominates the actual task of self-reflection, which suggests that it may not be a reliable method for estimating confidences.

4 External Evaluation

Judging others is typically easier than judging yourself. So in this section we experiment with using an LVLM as external evaluator to select which of the nine models provides the best answer to a given question. This is similar to the evaluators in [12, 23, 29]. As external evaluator we use a large, instruction tuned version of PaLM, denoted as `PaLM 2-L-IT`.

4.1 Choosing the Best Answer

A straightforward approach is to simply provide the external evaluator with all possible answers and have it select one. We try this approach. Furthermore, we also use more elaborate prompts with intermediate questions, for example to match the entity type of the question and the answer, or to emphasise answers which occur multiple times. Finally, we try including the retrieved Wikipedia context to aid the choice. However, preliminary experiments evaluating these three prompting strategies never achieve more than 43% accuracy, significantly lower than the 49.4% baseline given by PaLM with Lens (full results in Appendix F).

4.2 Evaluating the Reasoning Process

Since simple selection does not work, we resort to having the evaluator carefully analyse the reasoning process of each LVLM to judge whether the answer is sound or not.

A Lens-based LVLM (1) queries Lens to get the Wikipedia Entity and then (2) extracts the answer from its corresponding Wikipedia page. We have the evaluator analyse the success of both steps using multi-step prompts. In particular, for (1) we first ask whether Lens gave any result at all. If it did, we have the evaluator extract the entity type of the question (*e.g.* bird). Then we ask whether the retrieved entity (*e.g.* sparrow, Matterhorn) is indeed a bird or not. For (2) we give the evaluator the question and the Wikipedia page. Then we have it list all possible answers to the question, select the most informative answer (*e.g.* Central Europe is more informative than Europe), and compare that to the answer given by the Lens-based LVLM. If either of the two reasoning steps is judged unsuccessful, the answer of this LVLM is discarded. Full prompts are given in Appendix E.

For the PromptCap-based LVLMs we have a different strategy. PromptCap always gives a caption, so there are no failure modes there. Furthermore, the captions rarely explicitly contain the information

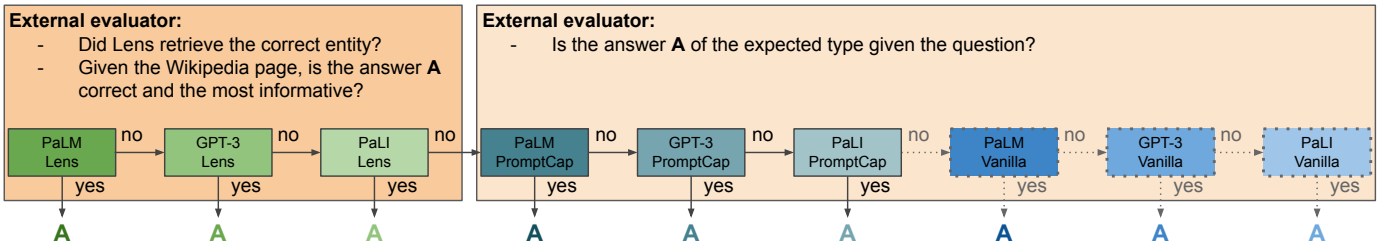

Figure 5: LVLM cascade by evaluating the reasoning process using an external evaluator.

Table 2: Results of cascade and evaluator performance.

	BEM
PaLM with Lens (baseline)	48.8%
Cascade with vanilla LVLMs	53.2%
Cascade without vanilla LVLMs	**53.4%**

(a) Accuracy of LVLM cascade.

	Precision	Recall
Lens	77.4%	76.7%
PromptCap	29.1%	86.3%
Vanilla	19.3%	86.5%

(b) Precision and recall of evaluator.

asked for in the question, so we cannot verify that either. Instead, we have the evaluator compare via a two-step prompt whether the answer of the LVLM is of the correct information type given the question: *e.g.* when asked about the name of a country we may expect Switzerland but not Europe. Since this strategy only relies on the question and the answer, we also use it for the vanilla LVLMs.

At this point, we have binary decisions for whether the answer for a single LVLM is sound. To combine everything into a single model, we use a cascade [9] which is visualized in Fig. 5. We choose the cascade order roughly based on performance: we first cascade all Lens-based LVLMs, then the PromptCap-based LVLMs, and then optionally the vanilla LVLMs. Within these, we first cascade PaLM, then GPT-3, then PaLI.

Results. Tab. 2a shows that we *finally* made a successful ensemble of our LVLMs: the cascade which excludes the vanilla LVLMs improves SOTA from from 48.8% to 53.4%; a total absolute improvement of 4.6%.

Analysis. To gain more insight into why our cascade works, we first look at how well our evaluator can judge the reasoning process. We do this in terms of precision (*i.e.* how many of the answers judged to be correct by the evaluator are truly correct) and recall (*i.e.* how many of the correct answers are also judged to be correct by the evaluator). This is shown in Tab. 2b. As can be seen, precision and recall are both high for evaluating the reasoning process of the Lens-based LVLMs, but precision is rather low for the other LVLMs. Hence our increased performance mostly stems from a good judgement of the Lens-based LVLM reasoning process.

Diving deeper into this result, we find that in the vast majority of the cases an answer is rejected because Lens did not retrieve any entity (!). This suggests a much easier cascade: Use the Lens-based LVLM if Lens retrieved an entity, otherwise resort to PromptCap to at least have some information of what is depicted. We implement this using PaLM leading to a simple 2-step cascade, which yields 52.4% BEM accuracy, close to the 53.4% of our best performing cascade. Hence most of our improvements simply come from observing whether Lens retrieved anything at all.

Finally, since our evaluator is based on the larger `PaLM 2-L-IT`, we also compare our cascade to the `PaLM 2-L-IT` + Lens model. This results in a BEM accuracy of 52.7%. Hence compared to this baseline, we achieve a much smaller performance improvement of 0.7%.

5 Conclusion

We investigated *how (not) to ensemble LVLMs for VQA* and found that an external evaluation model using the LVLMs in a cascade works best, increasing SOTA on EncyclopedicVQA from 48.8% to 53.4%. On the other hand, ablation on our cascade shows that a significant portion of the increase is given by (1) the larger size of our evaluator model, and (2) conditioning on whether Lens has identified a Wikipedia entity to decide which context should be added to the LVLM input.

There are additional limitations that derive from prompting LVLMs. Firstly, we only experimented with zero-shot prompts for extracting a confidence score via self-reflection. This is because, in our opinion, the inclusion of few-shot exemplars would risk biasing the model in this task. Additionally, the search space for natural language prompts is extremely large, and we do not claim to have carried out comprehensive prompt engineering.

Much of the large potential shown by the oracle experiment (a further 13.6% above the new SOTA) still remains untapped. Future research can look into strategies to capture more of it. Overall, we have tried a wide range of methods and found that improving performance on Encyclopedic-VQA via ensembling is difficult, despite the large gain promised by the oracle.

Acknowledgments and Disclosure of Funding

We thank Andre Araujo and Vitto Ferrari for their useful comments and insights.

References

[1] Google Lens. https://lens.google.com - Web interface available at https://images.google.com.

[2] J. B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: A visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022.

[3] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] J. Bulian, C. Buck, W. Gajewski, B. Börschinger, and T. Schuster. Tomayto, Tomahto. Beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[6] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. PaLM: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023.

[8] T. Dietterich. Ensemble learning. In *The Handbook of Brain Theory and Neural Networks, Second Edition*. The MIT Press, 2002.

[9] J. Gama and P. Brazdil. Cascade generalization. *Machine Learning*, 41:315–343, 12 2000.

[10] I. Gitman, V. Lavrukhin, A. Laptev, and B. Ginsburg. Confidence-based ensembles of end-to-end speech recognition models. In *INTERSPEECH*. International Speech Communication Association (ISCA), 2023.

[11] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[12] S. Hao, Y. Gu, H. Ma, J. J. Hong, Z. Wang, D. Z. Wang, and Z. Hu. Reasoning with language model is planning with world model. *ArXiv preprint*, 2023.

[13] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, and F. Wei. Language models are general-purpose interfaces. *ArXiv preprint*, 2022.

[14] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo. PromptCap: Prompt-guided task-aware image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2963–2975, 2023.

[15] S. Imani, A. Beyram, and H. Shrivastava. DiversiGATE: A comprehensive framework for reliable large language models. In *Workshop on Knowledge and Logical Reasoning in the Era of Data-driven Learning at ICML'23*, 2023.

[16] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] L. Li, Q. Hu, X. Wu, and D. Yu. Exploration of classification confidence in ensemble learning. *Pattern Recognition*, 47(9):3120–3131, 2014.

[18] T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *International Conference on Computer Vision (ICCV)*, 2023.

[19] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021.

[20] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, page 2901–2907. AAAI Press, 2015.

[21] D. Oniani, J. Hilsman, H. Dong, F. Gao, S. Verma, and Y. Wang. Large language models vote: Prompting for rare disease identification. *ArXiv preprint*, 2023.

[22] R. Rosales, P. Popov, and M. Paulitsch. Evaluation of confidence-based ensembling in deep learning image classification. *ArXiv preprint*, 2023.

[23] N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. *ArXiv preprint*, 2023.

[24] C. Si, Z. Gan, Z. Yang, S. Wang, J. Wang, J. Boyd-Graber, and L. Wang. Prompting GPT-3 to be reliable. In *International Conference on Learning Representations (ICLR)*, 2023.

[25] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *ArXiv preprint*, 2023.

[26] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. *ArXiv preprint*, 2023.

[27] Y. Yang, H. Lv, and N. Chen. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, 56:5545–5589, 2021.

[28] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022.

[29] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv preprint*, 2023.

A Oracle experiment results

In Tab. 3 we show a more extensive overview of the baseline performance and the performance of oracle ensembles.

Table 3: Oracle and Baseline results for EncyclopedicVQA. We show the results of the individual models (from [18]) and for different ensembles in the oracle setting. The three base models are PaLI, PaLM and GPT-3. The ensemble with all 9 models outperforms the best single model by 18.2%.

Ensemble	**Acc. (BEM)**
PaLI vanilla	13.0%
PaLI PromptCap	17.8%
PaLI Lens Section	28.1%
GPT-3 vanilla	15.5%
GPT-3 PromptCap	25.6%
GPT-3 Lens Section	44.6%
PaLM vanilla	19.7%
PaLM PromptCap	29.7%
PaLM Lens Section	***48.8%***
3 base models with PromptCap	38.9%
3 base models with Lens	56.6%
3 PaLM models (vanilla, PromptCap, Lens)	58.8%
6 base models (all PromptCap and all Lens)	65.8%
9 base models (all PromptCap, Lens, and vanilla)	**67.0%**

B Majority voting results

In Tab. 4 we show the results of majority voting using exact matching and BEM-based soft matching.

Table 4: Results of ensembling via majority voting. Exact matching is performed after normalisation via lower-casing and punctuation removal.

	Acc. (BEM)
PaLM with Lens (baseline)	48.8%
Majority voting via exact matching	40.9%
Majority voting via BEM soft-matching	45.3%

C Calibration

C.1 Re-calibration with temperature scaling

We use temperature scaling to re-calibrate the sequence probabilities for each of the LVLMs. For each model we select the temperature t^* with the lowest ECE / Brier score, simply by evaluating multiple temperature values.

ECE diagrams before and after temperature scaling are shown in Fig. 6.

C.2 Effect of length normalization on calibration

Normalizing the sequence probabilities for length results in worse ECE and Brier scores than using unnormalized probabilities. This is true for all the Lens-based LVLMs, as shown in Tab. 5. However, applying recalibration via temperature scaling significantly reduces this gap, and even gives (slightly)

better Brier scores for PaLI and GPT-3 after length normalization (average difference in ECE: 0.068±0.028 without recalibration, 0.015±0.019 with recalibration; average difference in Brier score: 0.031±0.023 without recalibration, 0.01±0.027 with recalibration). On the other hand, length normalization improves the BEM accuracy of all classical ensembling methods, as evidenced in Tab. 6. We hence use normalized probabilities.

C.3 Calibration in perspective

To put these calibration results in perspective, we compare our observations and results with the ImageNet classification experiments in [19]. In [19], 26 models trained on ImageNet-1K are evaluated for different calibration metrics. The authors observe correlation between the calibration metrics, and less of a correlation between calibration metrics and classification accuracy. Moreover, they report Brier scores in the range 17.6 - 58.2 (mean 29.9) and ECE scores in the range 1.4 - 8.4 (mean 3.7). The LVLMs on our task have ECE scores which are slightly higher, and Brier scores which are below their mean. Based on these results, we conclude that the sequence probabilities are (reasonably) well calibrated and we aim to compare these across the different models.

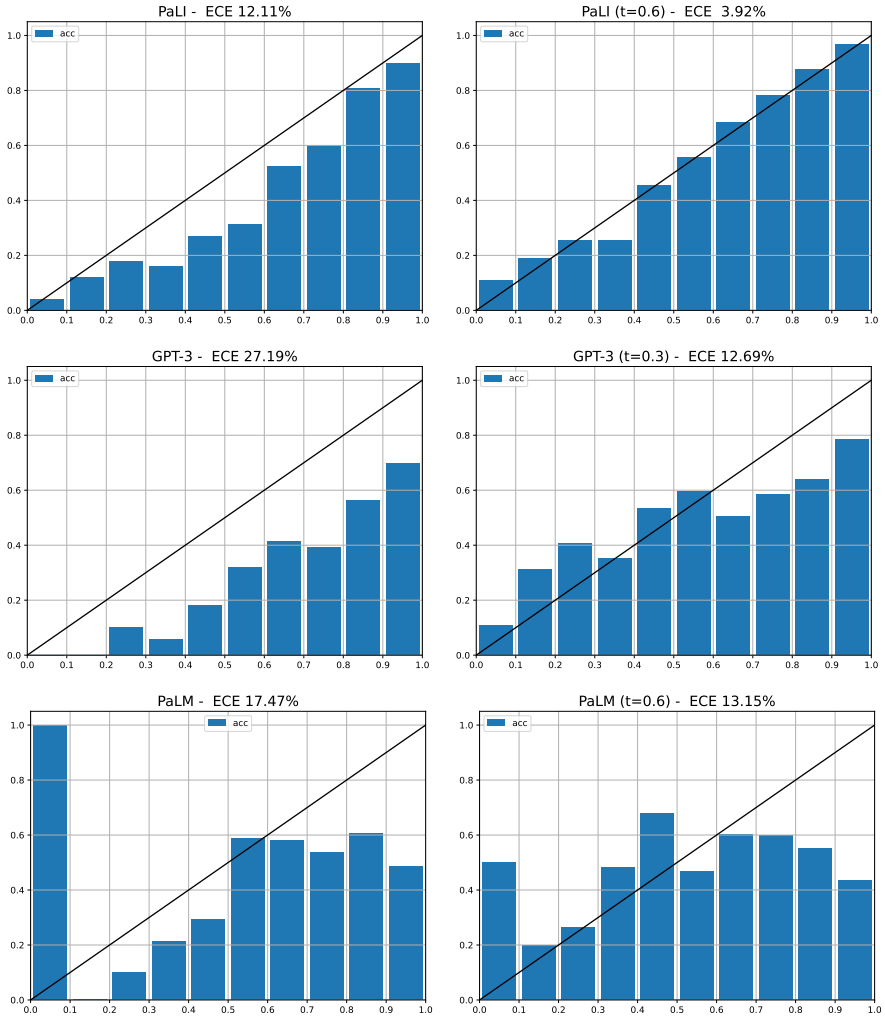

Figure 6: ECE diagrams for the three LVLMs before and after temperature scaling.

Table 5: Calibration results before and after normalizing the sequence probabilities for length.

	Before normalization			**After normalization**		
	PaLI	PaLM	GPT-3	PaLI	PaLM	GPT-3
ECE (↓)	0.081	0.107	0.177	0.121	0.175	0.272
Brier (↓)	0.174	0.219	0.246	0.182	0.272	0.278
ECE-re (↓)	0.037	0.095	0.121	0.039	0.131	0.127
Brier-re (↓)	0.167	0.215	0.229	0.164	0.256	0.221

Table 6: BEM accuracy of classical ensembling methods before and after normalizing the sequence probabilities for length. All results are on the Lens retrieval setup. Note that our logistic regression classification method one-vs-rest (OvR), taking the argmax of the three classifiers to output which LVLM to use.

	Before normalization	**After normalization**
Max probability	48.7%	49.2%
Weighted voting	48.6%	49.2%
Logistic regression	49.9%	50.6%

D Prompt Variations

In Fig. 7 we show the binary prompt variations for PaLI and PaLM for 10 different paraphrasings of the confidence prompt. We show the results for the three different base models (vanilla, PromptCap and Lens).

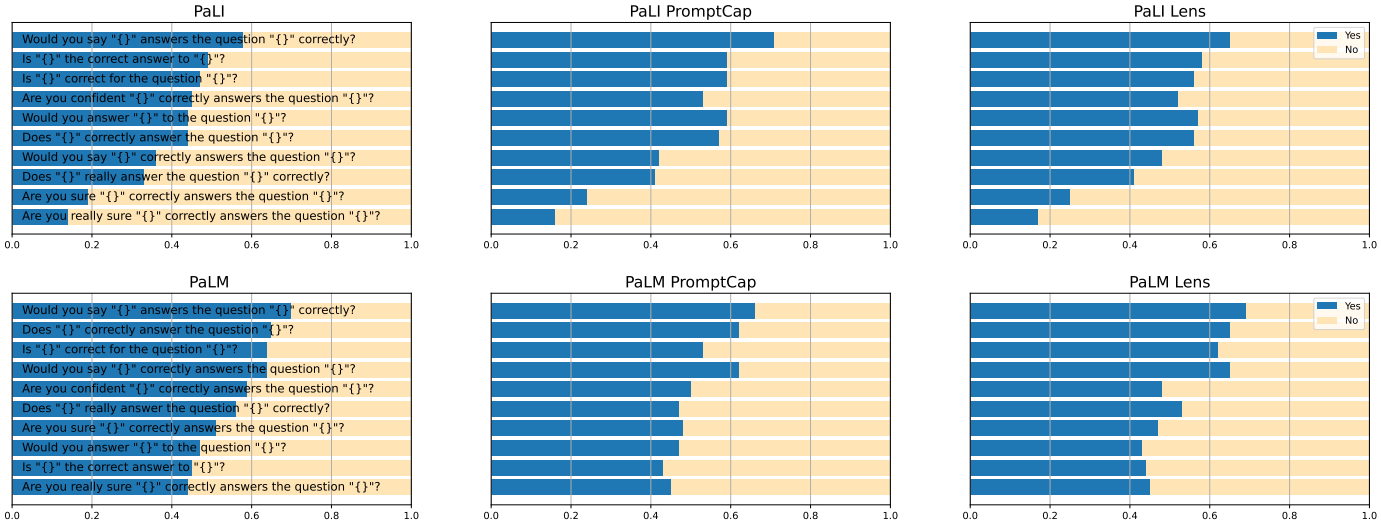

Figure 7: Prompt variations for PaLI and PaLM using the answers of three different base models. For each prompt the percentage of 'yes' answers is shown.

E Evaluator Model Prompts

All the exemplars in the following prompts are extracted from the Encyclopedic-VQA training set.

E.1 Prompts for evaluating Lens LVLMs

Prompt 1: Prompt to extract retrieval entity expected by the question.

```
[ENTITY TYPE]: What type of entity does the [QUESTION] refer to?

[QUESTION]: What is one common name for this plant?
[ENTITY TYPE]: plant

[QUESTION]: What ants engage in interference competition against
this ant colonies by 'plugging' entrances to the nest with sand and
small rocks?
[ENTITY TYPE]: ant

[QUESTION]: In what country is this amusement park located?
[ENTITY TYPE]: amusement park

[QUESTION]: What does this bird try to do to the dingo by jumping
into the air and kicking or stamping the dingo on its way down?
[ENTITY TYPE]: bird

[QUESTION]: What kind of shopping centre was there before this one?
[ENTITY TYPE]: shopping centre

[QUESTION]: How does this moth differ from the bride moths?
[ENTITY TYPE]: moth

[QUESTION]: During the 1995 anniversary of ve day, what did seven
heads of state take up at this hotel?
[ENTITY TYPE]: hotel

[QUESTION]: What kind of structure is this memorial?
[ENTITY TYPE]: memorial

[QUESTION]: What shape are the cones of this tree?
[ENTITY TYPE]: tree

[QUESTION]: What was the nationality of the architect who presented
plans for this theatre?
[ENTITY TYPE]: theatre

[QUESTION]: What is the habitat of this animal?
[ENTITY TYPE]: animal

[QUESTION]: Is this village accessible or inaccessible by road from
east bergholt?
[ENTITY TYPE]: village

[QUESTION]: What direction does the tunnel of this insect usually
go?
[ENTITY TYPE]: insect
```

```
[QUESTION]: What does this church do with its biblical foundations
and principles?
[ENTITY TYPE]: church

[QUESTION]: Besides being the oldest, what other distinction does
this pier have?
[ENTITY TYPE]: pier

[QUESTION]: What do adult this butterflies eat?
[ENTITY TYPE]: butterfly

[QUESTION]: What is depicted of king alexander jannaeus in this
model?
[ENTITY TYPE]: model

[QUESTION]: What sex of this dragonfly has a bright red abdomen?
[ENTITY TYPE]: dragonfly

[QUESTION]: What is the climate like in this city?
[ENTITY TYPE]: city

[QUESTION]: What does the queen of this wasp perform dance-like
wing of?
[ENTITY TYPE]: wasp

[QUESTION]: {}
[ENTITY TYPE]:
```

Prompt 2: Prompt to check retrieved entity matches expected entity.

```
Is the {} a {}?  Answer just yes or no.
```

Prompt 3: Prompt to extract most informative answer from retrieved Wikipedia context.

```
Answer the questions below.  Only show your answer to [MOST
INFORMATIVE ANSWER], do not show [ALL ANSWERS].

[ALL ANSWERS]: If there are more than one possible answer to the
[QUESTION] that can be found in the [EVIDENCE], give all possible
answers.  If no answers can be found in the [EVIDENCE], say 'no
answers'.

[MOST INFORMATIVE ANSWER]: Which answer is most informative
among the ones above?  Give that answer (hint:  informative means
containing the most precise and useful information).  If you
answered 'no answers' to [ALL ANSWERS], say 'no answer'.

[QUESTION]: {}
[EVIDENCE]: {}
[MOST INFORMATIVE ANSWER]:
```

Prompt 4: Prompt to assess equivalence of most informative answer and LVLM answer.

```
[EQUIVALENT OR NOT]: Given the [QUESTION], the [RESPONSE] and the
[MOST INFORMATIVE ANSWER], is the [RESPONSE] the same or equivalent
to the [MOST INFORMATIVE ANSWER]? Just say yes or no.

[QUESTION]: Where has this plant become a naturalized archaeophyte?
[RESPONSE]: Europe
[MOST INFORMATIVE ANSWER]: Central Europe
[EQUIVALENT OR NOT]: no

[QUESTION]: When do these birds defend fruit-bearing trees?
[RESPONSE]: summertime
[MOST INFORMATIVE ANSWER]: summer
[EQUIVALENT OR NOT]: yes

[QUESTION]: What other trees does this tree grow with?
[RESPONSE]: coniferous swamps
[MOST INFORMATIVE ANSWER]: coniferous swamps
[EQUIVALENT OR NOT]: yes

[QUESTION]: What might the aggressive reproductive characteristics
of this tree be beneficially exploited for?
[RESPONSE]: propagation
[MOST INFORMATIVE ANSWER]: biofuel
[EQUIVALENT OR NOT]: no

[QUESTION]: Is this tree a popular or unpopular choice for guitar
soundboards?
[RESPONSE]: Popular.
[MOST INFORMATIVE ANSWER]: popular
[EQUIVALENT OR NOT]: yes

[QUESTION]: When were the latest renovations or repairs in this
park?
[RESPONSE]: Recently
[MOST INFORMATIVE ANSWER]: 2022
[EQUIVALENT OR NOT]: no

[QUESTION]: Where does this bird depart from every march to april?
[RESPONSE]: migration
[MOST INFORMATIVE ANSWER]: North America
[EQUIVALENT OR NOT]: no

[QUESTION]: How long do the leaves of this plant last?
[RESPONSE]: year-round
[MOST INFORMATIVE ANSWER]: throughout the year
[EQUIVALENT OR NOT]: yes

[QUESTION]: When did henry ii begin spending money on this castle?
[RESPONSE]: 1173
[MOST INFORMATIVE ANSWER]: 1176
[EQUIVALENT OR NOT]: no

[QUESTION]: Along with spain, where is this bird found?
[RESPONSE]: australia and new zealand.
[MOST INFORMATIVE ANSWER]: New Zealand.
[EQUIVALENT OR NOT]: yes
```

```
[QUESTION]: In meters, how long is this cathedral?
[RESPONSE]: 120 m
[MOST INFORMATIVE ANSWER]: 120
[EQUIVALENT OR NOT]: yes

[QUESTION]: How many millimeters long are the glumes of this plant?
[RESPONSE]: 2mm
[MOST INFORMATIVE ANSWER]: 1.5-2.5
[EQUIVALENT OR NOT]: yes

[QUESTION]: What is another predator of this ant besides insects
and birds?
[RESPONSE]: Mammals.
[MOST INFORMATIVE ANSWER]: tarantulas
[EQUIVALENT OR NOT]: no

[QUESTION]: When do this plant's flowers form fruit?
[RESPONSE]: after pollination
[MOST INFORMATIVE ANSWER]: after flowering
[EQUIVALENT OR NOT]: no

[QUESTION]: When did hofwart de kirchheim reside at this castle?
[RESPONSE]: 1200s
[MOST INFORMATIVE ANSWER]: 13th century
[EQUIVALENT OR NOT]: yes

[QUESTION]: When was the first record of this plant in ireland?
[RESPONSE]: 11th century
[MOST INFORMATIVE ANSWER]: 12th century
[EQUIVALENT OR NOT]: no

[QUESTION]: What is the traditional use of this plant in medicine?
[RESPONSE]: to stop bleeding.
[MOST INFORMATIVE ANSWER]: stop blood loss
[EQUIVALENT OR NOT]: yes

[QUESTION]: In addition to deltas and marshes, where does this bird
feed?
[RESPONSE]: estuaries.
[MOST INFORMATIVE ANSWER]: wetlands
[EQUIVALENT OR NOT]: no

[QUESTION]: What does this castle display about Richard Pennant?
[RESPONSE]: Owned nearly 1,000 enslaved people.
[MOST INFORMATIVE ANSWER]: slavery links
[EQUIVALENT OR NOT]: yes

[QUESTION]: How often were rockets fired from this lighthouse
starting in 1896?
[RESPONSE]: Every year
[MOST INFORMATIVE ANSWER]: annually
[EQUIVALENT OR NOT]: yes

[QUESTION]: {}
[RESPONSE]: {}
[MOST INFORMATIVE ANSWER]: {}
[EQUIVALENT OR NOT]:
```

E.2 Prompts for evaluating PromptCap and Vanilla LVLMs

Prompt 5: Prompt to extract the type of information required by the question.

```
[REQUIRED INFORMATION]: What information does the [QUESTION] ask
for?

[QUESTION]: What is one common name for this plant?
[REQUIRED INFORMATION]: common plant name

[QUESTION]: What ants engage in interference competition against
this ant colonies by 'plugging' entrances to the nest with sand and
small rocks?
[REQUIRED INFORMATION]: ant species

[QUESTION]: In what country is this amusement park located?
[REQUIRED INFORMATION]: name of country

[QUESTION]: What does this bird try to do to the dingo by jumping
into the air and kicking or stamping the dingo on its way down?
[REQUIRED INFORMATION]: purpose or action directed to the dingo

[QUESTION]: What kind of shopping centre was there before this one?
[REQUIRED INFORMATION]: kind of shopping centre

[QUESTION]: How does this moth differ from the bride moths?
[REQUIRED INFORMATION]: habit or feature a moth has

[QUESTION]: During the 1995 anniversary of ve day, what did seven
heads of state take up at this hotel?
[REQUIRED INFORMATION]: action taken up at hotel

[QUESTION]: What kind of structure is this memorial?
[REQUIRED INFORMATION]: kind of structure

[QUESTION]: What shape are the cones of this tree?
[REQUIRED INFORMATION]: shape

[QUESTION]: What was the nationality of the architect who presented
plans for this theatre?
[REQUIRED INFORMATION]: nationality

[QUESTION]: What is the habitat of this animal?
[REQUIRED INFORMATION]: habitat of the animal

[QUESTION]: Is this village accessible or inaccessible by road from
east bergholt?
[REQUIRED INFORMATION]: yes/no answer

[QUESTION]: What direction does the tunnel of this insect usually
go?
[REQUIRED INFORMATION]: direction

[QUESTION]: What does this church do with its biblical foundations
and principles?
[REQUIRED INFORMATION]: action taken by church
```

```
[QUESTION]: Besides being the oldest, what other distinction does
this pier have?
[REQUIRED INFORMATION]: unique feature of pier

[QUESTION]: What do adult this butterflies eat?
[REQUIRED INFORMATION]: butterfly's prey or food

[QUESTION]: What is depicted of king alexander jannaeus in this
model?
[REQUIRED INFORMATION]: property or belonging

[QUESTION]: What sex of this dragonfly has a bright red abdomen?
[REQUIRED INFORMATION]: sex of the animal

[QUESTION]: What is the climate like in this city?
[REQUIRED INFORMATION]: type of climate

[QUESTION]: When was this landmark built?
[REQUIRED INFORMATION]: year of construction

[QUESTION]: {}
[REQUIRED INFORMATION]:
```

Prompt 6: Prompt to determine if the answer contains the required information.

```
[GIVES REQUIRED INFORMATION]: Does the [ANSWER] to the [QUESTION]
satisfy the [REQUIRED INFORMATION]?

[QUESTION]: What is one common name for this plant?
[REQUIRED INFORMATION]: common plant name
[ANSWER]: Peacock Flower
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What ants engage in interference competition against
this ant colonies by 'plugging' entrances to the nest with sand and
small rocks?
[REQUIRED INFORMATION]: ant species
[ANSWER]: ants
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: In what country is this amusement park located?
[REQUIRED INFORMATION]: name of country
[ANSWER]: austria
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What does this bird try to do to the dingo by jumping
into the air and kicking or stamping the dingo on its way down?
[REQUIRED INFORMATION]: purpose or action directed to the dingo
[ANSWER]: scare it away
[GIVES REQUIRED INFORMATION]: yes
```

```
 [QUESTION]: What kind of shopping centre was there before this
one?
[REQUIRED INFORMATION]: kind of shopping centre
[ANSWER]: Westfield shopping centre
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: How does this moth differ from the bride moths?
[REQUIRED INFORMATION]: habit or feature a moth has
[ANSWER]: diurnal
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: During the 1995 anniversary of ve day, what did seven
heads of state take up at this hotel?
[REQUIRED INFORMATION]: action taken up at hotel
[ANSWER]: Conrad London
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What kind of structure is this memorial?
[REQUIRED INFORMATION]: kind of structure
[ANSWER]: obelisk
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What shape are the cones of this tree?
[REQUIRED INFORMATION]: shape
[ANSWER]: leaves and tendrils
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What was the nationality of the architect who presented
plans for this theatre?
[REQUIRED INFORMATION]: nationality
[ANSWER]: Le Corbusier
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What is the habitat of this animal?
[REQUIRED INFORMATION]: habitat of the animal
[ANSWER]: river deltas
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: Is this village accessible or inaccessible by road from
east bergholt?
[REQUIRED INFORMATION]: yes/no answer
[ANSWER]:
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What direction does the tunnel of this insect usually
go?
[REQUIRED INFORMATION]: direction
[ANSWER]: termites
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What does this church do with its biblical foundations
and principles?
[REQUIRED INFORMATION]: action taken by church
[ANSWER]: people praying
[GIVES REQUIRED INFORMATION]: no
```

```
 [QUESTION]: Besides being the oldest, what other distinction does
this pier have?
[ENTITY TYPE]: unique feature of pier
[ANSWER]: longest
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What do adult this butterflies eat?
[REQUIRED INFORMATION]: butterfly's prey or food
[ANSWER]: pollen
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What is depicted of king alexander jannaeus in this
model?
[REQUIRED INFORMATION]: human feature, property or belonging
[ANSWER]: Jerusalem model
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: What sex of this dragonfly has a bright red abdomen?
[REQUIRED INFORMATION]: sex
[ANSWER]: male
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: What is the climate like in this city?
[REQUIRED INFORMATION]: type of climate
[ANSWER]: New York
[GIVES REQUIRED INFORMATION]: no

[QUESTION]: When was this landmark built?
[ENTITY TYPE]: year of construction
[ANSWER]: 1790
[GIVES REQUIRED INFORMATION]: yes

[QUESTION]: {}
[REQUIRED INFORMATION]: {}
[ANSWER]: {}
[GIVES REQUIRED INFORMATION]:
```

F Answer choice results

In Tab. 7 we show the results of choosing the best answer among those given by the LVLMs, using different prompts.

Table 7: Results of choosing the best answer via prompting, all well below the baseline (48.8%). When `[EXEMPLARS]` are part of the prompt, these are the same as in Appendix E.1 Prompt 1.

Sample size	Prompt	Acc. (BEM)

100	`Consider the question [QUESTION] and the list with possible answers [ALL ANSWERS].`
	`[FINAL ANSWER]: Among the answers in [ALL ANSWERS], which one is the most likely to be correct?`
	`Hint: Repeated answers are more likely to be correct.`
		40.0%

	`Consider the question [QUESTION], the answer [ANSWER] and the list with possible answers [POSSIBLE ANSWERS].`
	`[SIMILAR ANSWERS]: Count the number of answers in [POSSIBLE ANSWERS] which are identical or semantically similar to [ANSWER].`
	`If the majority of the answers are similar, then [ANSWER] is the [CORRECT ANSWER].`
	`If not, consider the following:`
	`[BETTER ANSWER]: Is there an answer in [POSSIBLE ANSWERS] which answers the question [QUESTION] more precisely and concisely than [ANSWER]?`
	`If yes, then this answer is the [CORRECT ANSWER].`
	`If not, use [ANSWER] as the [CORRECT ANSWER].`
		43.0%

500	`(1)`
	`[ENTITY TYPE]: Based on the [QUESTION], what kind of entity would you expect the answer to be?`
	`[EXEMPLARS]`
	
	`(2)`
	`[BEST ANSWER]: Given the [QUESTION], the comma-separated [ANSWERS] and the [ENTITY TYPE], which of the [ANSWERS] is the best and most informative for the question while also matching the [ENTITY TYPE]? You must choose only one of the comma-separated [ANSWERS] and output it without changes.`
		40.0%

	`(1)`
	`Answer the questions below. Only show your answer to [MOST INFORMATIVE ANSWER], do not show [ALL ANSWERS].`
	`[ALL ANSWERS]: If there are more than one possible answer to the [QUESTION] that can be found in the [EVIDENCE], give all possible answers. If no answer can be found in the [EVIDENCE], say 'no answers'.`
	`[MOST INFORMATIVE ANSWER]: Which answer is most informative among the ones above? Give that answer (hint: informative means containing the most precise and useful information).If you answered 'no answers' to [ALL ANSWERS], output 'no answer'.`
	
	`(2)`
	`[CHOICE]: Given the [QUESTION], the comma-separated [RESPONSES] and the [MOST INFORMATIVE ANSWER], which of the [RESPONSES] is most equivalent to the [MOST INFORMATIVE ANSWER]? Only return that response.`
		32.6%