Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering

Zhenwei Shao[1] Zhou Yu[1*] Meng Wang[2] Jun Yu[1]

[1]Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, China.
[2]School of Computer Science and Information Engineering, Hefei University of Technology, China

A. Broader Impact

From a broader "VL+LLM" point of view, Prophet connects a vision-language (VL) model and a frozen LLM, aiming to endow the VL model with knowledge reasoning ability. Compared with Flamingo [1] which learns a large *general* VL model on top of a frozen LLM (Chinchilla-70B) in an end-to-end manner, Prophet introduces a decoupled VL+LLM learning paradigm to learn a small *task-specific* VL model independently and then feed its predictions to any external LLM for refinement. Consequently, Prophet significantly reduces the computational costs of model training and also benefits from the flexible choices of LLMs (both the offline and online LLMs are supported). We hope the decoupled learning paradigm in Prophet will inspire future research using LLMs.

B. Implementation Details

B.1. Improved MCAN Model

Model architecture. We propose an improved version of MCAN [11] based on its open-sourced MCAN-large implementation. Our modifications to the model architecture include: (i) we replace the original bottom-up-attention features with the grid-based features extracted from the CLIP's visual encoder with RN50×64 backbone [6]; (ii) we introduce the RoPE mechanism [7] to each image self-attention layer of MCAN to supplement the grid-based features with positional information; and (iii) we replace the original LSTM network with a pre-trained BERT-large model [2] as the text encoder before MCAN. Table 1 shows the accuracies of different model variants on the testing set of OK-VQA. By progressively adding the modifications to the original MCAN model, our improved MCAN model reports a 53.0% accuracy, which is on par with current state-of-the-art methods like KAT [4].

Training recipe. We first pretrain the model on the augmented *train+val+vg* dataset from VQAv2 [3] and Visual Genome [5], with excluding the samples whose images are used in the testing split of OK-VQA to avoid data contamination. The settings for the pretraining stage are identical to the original implementation of MCAN. After that, the model is finetuned on the downstream OK-VQA and A-OKVQA datasets, respectively. For finetuning, the commonly used strategy is to replace the last linear layer (*i.e.*, the classification layer) with a new layer to adapt to the answer vocabulary of the downstream dataset. However, the answer vocabularies of the pretraining and finetuning datasets are *partially* overlapped. To maximally utilize the pretrained model parameters in the last layer, we inherit the parameters of existing answers and append new parameters for the new answers. After that, we freeze all the pretrained parameters and only update the new parameters for one epoch as a warm-up, and then train all model parameters for the rest training epochs. The detailed settings for the finetuning stage are shown in Table 2.

case	OK-VQA accuracy
original MCAN	43.6
+ CLIP visual feats	49.6
+ RoPE mechanism	50.3
+ BERT as the text encoder	53.0

Table 1. **Ablations for model architectures.** '+' denotes each modification is appended to the variant on the previous row.

Table 3 shows the effects of different training strategies. Even without finetuning, the pretrained model (b) is superior to the model trained from scratch (a), implying the importance of pretraining. Moreover, our new finetuning strategy (d) leads to significantly better performance than the commonly used strategy (c), showing the effectiveness of inheriting model parameters for existing answers.

B.2. Prompt Formats

We show an exemplar of Prophet's prompt in Table 7 and an exemplar of Prophet-MC's prompt (only used in the multiple-choice task of A-OKVQA) in Table 8.

[*]Zhou Yu is the corresponding author

config	setting
optimizer	AdamW
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.98
batch size	64
warm-up learning rate	1e-3
warm-up strategy	only update new parameters
warm-up epochs	1
base learning rate	5e-5
learning rate schedule	step decay
learning rate decay rate	0.2
learning rate decay epoch	6
total training epochs	6

Table 2. **Training settings.** These hyper-parameters are used in both the OK-VQA and A-OKVQA experiments.

training strategy	OK-VQA accuracy
(a) train from scratch	35.6
(b) pretrain, w/o finetune	41.1
(c) w/ finetune, replace last layer	47.7
(d) w/ finetune, append new answers	**53.0**

Table 3. **Ablations for training strategies.** Four variants use the same architecture (last row in Table 1) except for the last layer.

B.3. Choice of Other Open-source LLMs

We have tried to replace GPT-3 with another open-source LLM GPT-J (6B) [8] and observed a distinct accuracy drop (53.6% *vs.* 61.1%). This can be explained that GPT-J's in-context learning ability is not strong enough to understand the answer candidates. As reported in [9], this advanced ability *emerges* only when a LLM reaches a certain model size (*e.g*, >100B parameters). As running such a LLM offline is far beyond the computational resource we have, using GPT-3 is our best choice at this moment. If a proper-size and open-source LLM with the similar in-context learning ability to GPT-3 is available in the future, we can use it to replace GPT-3 seamlessly.

C. More Statistical Analyses

We provide more in-depth analyses of Prophet's performance on the testing set of OKVQA. All results are carried out using the default settings.

First, we show the per-type accuracies of MCAN (stage-1) and Prophet (stage-2) in Table 4. Prophet outperforms MCAN on all categories, indicating that generality of the knowledge in GPT-3. The improvement on the "Science and Technology" category is not as large as the rest categories. which can be explained that the required knowledge for this category is more specialized and professional. These questions are also challenging for humans.

Then, we display the distribution of four situations of Prophet's predictions before and after GPT-3 in Table 5.

category	MCAN	Prophet
Plants and Animals	52.58	**63.67**
Science and Technology	48.10	**48.81**
Sports and Recreation	59.08	**66.00**
Geography, History, Language and Culture	52.48	**62.98**
Brands, Companies and Products	51.98	**54.77**
Vehicles and Transportation	50.82	**58.01**
Cooking and Food	55.53	**62.09**
Weather and Climate	65.12	**68.37**
People and Everyday life	49.44	**54.67**
Objects, Material and Clothing	50.05	**57.20**

Table 4. **Per-category accuracies** of MCAN (stage-1) and Prophet (stage-2). This performance improvements of using GPT-3 are observed on all categories.

after before	correct	wrong
correct	54.4%	4.2%
wrong	12.0%	29.4%

Table 5. **The distribution of four situations** of Prophet's predictions before and after GPT-3. Prophet maintains the majority of correct predictions by MCAN, and the accuracy improvement by GPT-3 is mainly because the number of *wrong-to-correct* samples is larger than that of the *correct-to-wrong* samples.

failure cause	proportion
(a) insufficient visual understanding	27.3%
(b) incorrect knowledge reasoning	**44.1%**
(c) correct but differently expressed answer	22.8%
(d) others	5.8%

Table 6. **The distribution of failure causes** by human studies.

From the results, we draw the following conclusions: (i) The proportion of the *correct-to-correct* samples (54.4%) is close to the accuracy of MCAN (53.0%) and proportion of the *correct-to-wrong* samples (4.2%) are relatively small. This means that Prophet maintains the majority of correct predictions; (ii) the accuracy improvement of Prophet is mainly due to the fact that the proportion of *wrong-to-correct* samples (12.4%) is larger than that of the *correct-to-wrong* samples (4.2%); (iii) there are still a considerable amount of samples (29.4%) that both MCAN and Prophet fail to give the correct answer, which leaves sufficient room for future improvement.

Finally, we perform human studies to analyze the causes of wrong predictions in Table 6. For each category, we randomly sample 10% testing samples that Prophet fails to get the correct answer. This results in 172 samples. We ask three annotators to categorize each sample into one of the following four failure causes: (a) insufficient visual understanding; (b) incorrect knowledge reasoning; (c) correct but differently expressed answer; (d) others (*e.g.*, the failure is caused by the ambiguity of the question). From the results, we can see that the cause of "(b) incorrect knowledge reasoning" accounts for the highest proportion, which suggests that the bottleneck of Prophet still lies in the knowledge acquisition and reasoning. The cause of "(a) insufficient visual understanding" has the second highest proportion, showing the potential of devising more powerful VQA models. The cause of "(c) correct but differently expressed answer" also accounts for a considerable proportion. This reflects the limitation of the annotations and evaluation metric of OK-VQA.

D. Case Study

Figure 1 provides some testing samples along with their in-context examples to illustrate how the answer heuristics work. The results show that both the answer candidates and the answer-aware examples are helpful for GPT-3 to generate high-quality answers. It is worth noting that even though some predicted answers do not hit the ground-truth answers, they are still reasonable under human evaluation.

Figure 2 demonstrates some testing samples from different knowledge categories. In the 1st-3rd columns, we show the correctly answered samples with different prediction behaviors (*i.e.*, keep top-1, in top 2-K, and beyond top-K). The visualized results indicate that Prophet can adaptively choose suitable answers from candidates. In the last column, we show some failure samples, implying that there is still room for future improvement.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 1

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 1

[4] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL*, 2021. 1, 4

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[7] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 1

[8] Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021. 2

[9] Jason Wei, Yi Tay, Rishi Bommasani, et al. Emergent abilities of large language models. *TMLR*, 2023. 2

[10] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, pages 3081–3089, 2022. 4

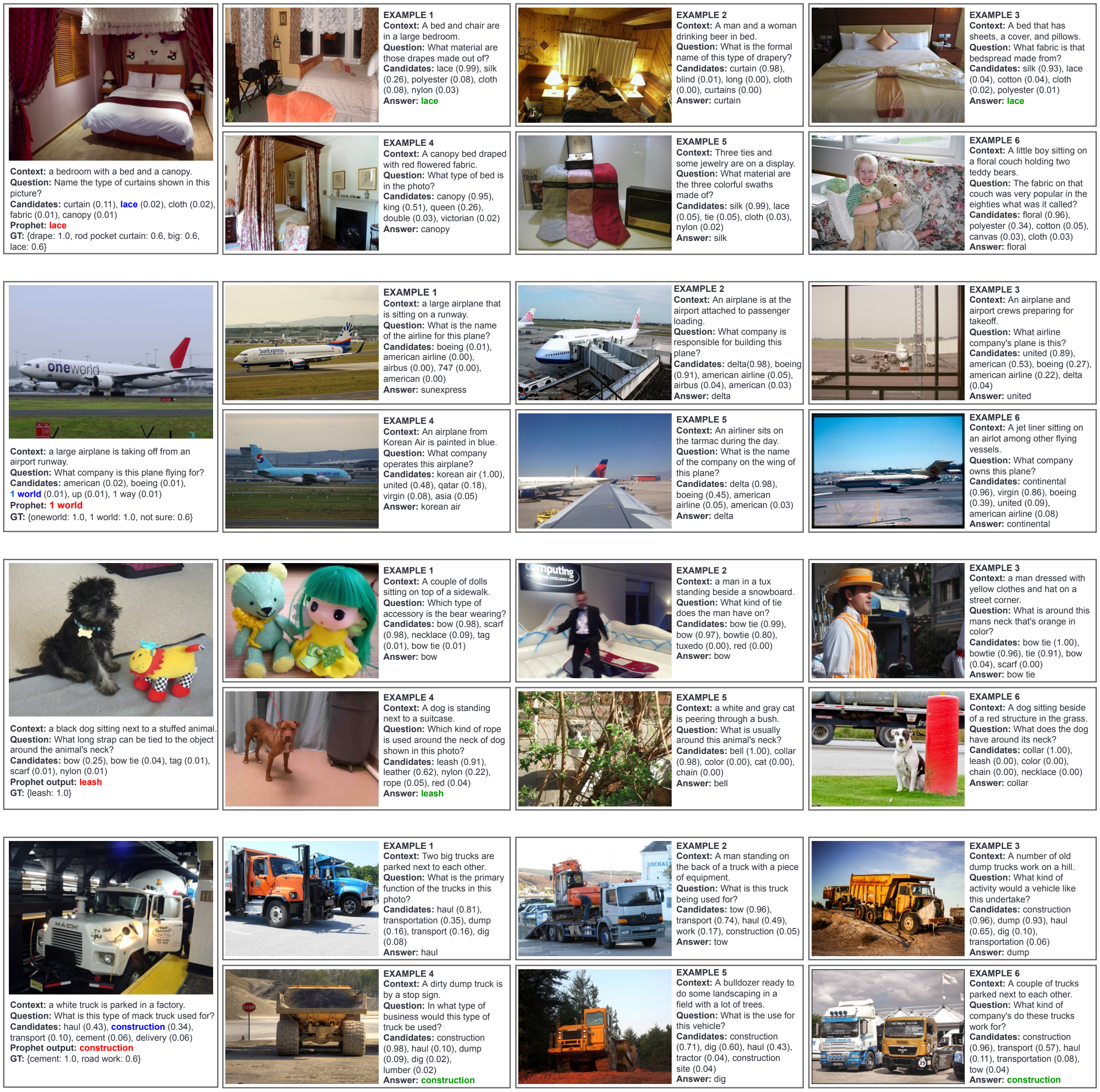
[11] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019. 1

```
Please answer the question according to the context and the answer candidates. Each answer
candidate is associated with a confidence score within a bracket. The true answer may not be
included in the candidates.

===
Context: The motorcycle racers are getting ready for a race.
===
Question: What sport are these guys doing?
===
Candidates: motorcross(0.94), motocross(0.79), bike(0.35), dirt bike(0.28), motorcycle(0.03),
bmx(0.03), cycling(0.02), motorbike(0.02), race(0.02), bicycle(0.02)
===
Answer: motorcross

===
Context: A man riding on the back of a motorcycle.
===
Question: Can you name a sport this person could be a
part of?
===
Candidates: race(0.94), motocross(0.80), dirt bike(0.70), motorcross(0.25), motorcycle(0.08),
bike(0.03), cycling(0.03), motorbike(0.01), ride(0.01), bicycle(0.01)
===
Answer: race

===
Context: A man on a motorcycle flying through the air.
===
Question: Which sport is this?
===
Candidates: motorcross(0.91), motocross(0.91), dirt bike(0.25), bike(0.05), bmx(0.03),
motorbike(0.01), motorcycle(0.01), bicycling(0.01), cycling(0.01), dirt(0.01)
===
Answer: motocross

===
Context: a black motorcycle parked in a parking lot.
===
Question: What sport can you use this for?
===
Candidates: race(0.53), motorcycle(0.41), motocross(0.19), bike(0.17), motorcross(0.15),
cycling(0.11), dirt bike(0.10), ride(0.08), bicycling(0.01), bicycle(0.01)
===
Answer:
```

Table 7. **A prompt exemplar for Prophet.** We only show 3 in-context examples here for better visualization. Following the implementations in PICa [10] and KAT [4], we use a special symbol '===' to separate each two lines.

```
Please choose the correct answer in the choices according to the context, the question and the
answer candidates. Each answer candidate is associated with a confidence score within a bracket.
The true answer may not be included in the candidates.

===
Context: A young man riding a skateboard on a sidewalk.
===
Question: What part of his body will be most harmed by the item in his mouth?
===
Candidates: skateboard(0.02), nothing(0.02), table(0.01), leg(0.01), helmet(0.00), knees(0.00),
skateboarding(0.00), head(0.00), teeth(0.00), falling(0.00)
===
Choices: (A) back, (B) lungs, (C) feet, (D) eyes
===
Answer: (B)

===
Context: A guy jumping in the air on a skateboard.
===
Question: What is touching the skateboard?
===
Candidates: shoe(0.89), foot(0.67), shoes(0.51), feet(0.45), sneakers(0.17), boy(0.03),
road(0.03), man(0.03), skateboard(0.02), nothing(0.02)
===
Choices: (A) sneakers, (B) dress shoes, (C) pogo stick, (D) hands
===
Answer: (A)

===
Context: A man flying through the air over a skateboard.
===
Question: Before going aloft what did the man ride?
===
Candidates: skateboard(1.00), skating(0.03), skate board(0.01), skate(0.01), board(0.01),
skateboarding(0.00), scooter(0.00), car(0.00), trick(0.00), longboard(0.00)
===
Choices: (A) unicycle, (B) skateboard, (C) plane, (D) car
===
Answer: (B)

===
Context: a young boy kneeling on a skateboard on the street.
===
Question: What did this lad likely injure here?
===
Candidates: skateboard(0.18), shoes(0.02), shoe(0.02), skateboarding(0.01), street(0.01),
flowers(0.01), skating(0.01), boy(0.01), head(0.00), skateboarder(0.00)
===
Choices: (A) knee, (B) elbow, (C) rear, (D) board
===
Answer:
```

Table 8. **A prompt exemplar for Prophet-MC.** Compared to the prompt in Table 7, we add one extra line of choices for the example and testing input, and change the output format to adapt to the multiple-choice task. All the differences are marked in red.



Context: a bedroom with a bed and a canopy.
Question: Name the type of curtains shown in this picture?
Candidates: curtain (0.11), lace (0.02), cloth (0.02), fabric (0.01), canopy (0.01)
Prophet: lace
GT: {drape: 1.0, rod pocket curtain: 0.6, big: 0.6, lace: 0.6}

EXAMPLE 1
Context: A bed and chair are in a large bedroom.
Question: What material are those drapes made out of?
Candidates: lace (0.99), silk (0.26), polyester (0.08), cloth (0.08), nylon (0.03)
Answer: lace

EXAMPLE 2
Context: A man and a woman drinking beer in bed.
Question: What is the formal name of this type of drapery?
Candidates: curtain (0.98), blind (0.01), long (0.00), cloth (0.00), curtains (0.00)
Answer: curtain

EXAMPLE 3
Context: A bed that has sheets, a cover, and pillows.
Question: What fabric is that bedspread made from?
Candidates: silk (0.93), lace (0.04), cotton (0.04), cloth (0.02), polyester (0.01)
Answer: lace

EXAMPLE 4
Context: A canopy bed draped with red flowered fabric.
Question: What type of bed is in the photo?
Candidates: canopy (0.95), king (0.51), queen (0.26), double (0.03), victorian (0.02)
Answer: canopy

EXAMPLE 5
Context: Three ties and some jewelry are on a display.
Question: What material are the three colorful swaths made of?
Candidates: silk (0.99), lace (0.05), tie (0.05), cloth (0.03), nylon (0.02)
Answer: silk

EXAMPLE 6
Context: A little boy sitting on a floral couch holding two teddy bears.
Question: The fabric on that couch was very popular in the eighties what was it called?
Candidates: floral (0.96), polyester (0.34), cotton (0.05), canvas (0.03), cloth (0.03)
Answer: floral

Context: a large airplane is taking off from an airport runway.
Question: What company is this plane flying for?
Candidates: american (0.02), boeing (0.01), 1 world (0.01), up (0.01), 1 way (0.01)
Prophet: 1 world
GT: {oneworld: 1.0, 1 world: 1.0, not sure: 0.6}

EXAMPLE 1
Context: a large airplane that is sitting on a runway.
Question: What is the name of the airline for this plane?
Candidates: boeing (0.01), american airline (0.00), airbus (0.00), 747 (0.00), american (0.00)
Answer: sunexpress

EXAMPLE 2
Context: An airplane is at the airport attached to passenger loading.
Question: What company is responsible for building this plane?
Candidates: delta(0.98), boeing (0.91), american airline (0.05), airbus (0.04), american (0.03)
Answer: delta

EXAMPLE 3
Context: An airplane and airport crews preparing for takeoff.
Question: What airline company's plane is this?
Candidates: united (0.89), american (0.53), boeing (0.27), american airline (0.22), delta (0.04)
Answer: united

EXAMPLE 4
Context: An airplane from Korean Air is painted in blue.
Question: What company operates this airplane?
Candidates: korean air (1.00), united (0.48), qatar (0.18), virgin (0.08), asia (0.05)
Answer: korean air

EXAMPLE 5
Context: An airliner sits on the tarmac during the day.
Question: What is the name of the company on the wing of this plane?
Candidates: delta (0.98), boeing (0.45), american airline (0.05), american (0.03)
Answer: delta

EXAMPLE 6
Context: A jet liner sitting on an airlot among other flying vessels.
Question: What company owns this plane?
Candidates: continental (0.96), virgin (0.86), boeing (0.39), united (0.09), american airline (0.08)
Answer: continental

Context: a black dog sitting next to a stuffed animal.
Question: What long strap can be tied to the object around the animal's neck?
Candidates: bow (0.25), bow tie (0.04), tag (0.01), scarf (0.01), nylon (0.01)
Prophet output: leash
GT: {leash: 1.0}

EXAMPLE 1
Context: A couple of dolls sitting on top of a sidewalk.
Question: Which type of accessory is the bear wearing?
Candidates: bow (0.98), scarf (0.98), necklace (0.09), tag (0.01), bow tie (0.01)
Answer: bow

EXAMPLE 2
Context: a man in a tux standing beside a snowboard.
Question: What kind of tie does the man have on?
Candidates: bow tie (0.99), bow (0.97), bowtie (0.80), tuxedo (0.00), red (0.00)
Answer: bow

EXAMPLE 3
Context: a man dressed with yellow clothes and hat on a street corner.
Question: What is around this mans neck that's orange in color?
Candidates: bow tie (1.00), bowtie (0.96), tie (0.91), bow (0.04), scarf (0.00)
Answer: bow tie

EXAMPLE 4
Context: A dog is standing next to a suitcase.
Question: Which kind of rope is used around the neck of dog shown in this photo?
Candidates: leash (0.91), leather (0.62), nylon (0.22), rope (0.05), red (0.04)
Answer: leash

EXAMPLE 5
Context: a white and gray cat is peering through a bush.
Question: What is usually around this animal's neck?
Candidates: bell (1.00), collar (0.98), color (0.00), cat (0.00), chain (0.00)
Answer: bell

EXAMPLE 6
Context: A dog sitting beside of a red structure in the grass.
Question: What does the dog have around its neck?
Candidates: collar (1.00), leash (0.00), color (0.00), chain (0.00), necklace (0.00)
Answer: collar

Context: a white truck is parked in a factory.
Question: What is this type of mack truck used for?
Candidates: haul (0.43), construction (0.34), transport (0.10), cement (0.06), delivery (0.06)
Prophet output: construction
GT: {cement: 1.0, road work: 0.6}

EXAMPLE 1
Context: Two big trucks are parked next to each other.
Question: What is the primary function of the trucks in this photo?
Candidates: haul (0.81), transportation (0.35), dump (0.16), transport (0.16), dig (0.08)
Answer: haul

EXAMPLE 2
Context: A man standing on the back of a truck with a piece of equipment.
Question: What is this truck being used for?
Candidates: tow (0.96), transport (0.74), haul (0.49), work (0.17), construction (0.05)
Answer: tow

EXAMPLE 3
Context: A number of old dump trucks work on a hill.
Question: What kind of activity would a vehicle like this undertake?
Candidates: construction (0.96), dump (0.93), haul (0.65), dig (0.10), transportation (0.06)
Answer: dump

EXAMPLE 4
Context: A dirty dump truck is by a stop sign.
Question: In what type of business would this type of truck be used?
Candidates: construction (0.98), haul (0.10), dump (0.09), dig (0.02), lumber (0.02)
Answer: construction

EXAMPLE 5
Context: A bulldozer ready to do some landscaping in a field with a lot of trees.
Question: What is the use for this vehicle?
Candidates: construction (0.71), dig (0.60), haul (0.43), tractor (0.04), construction site (0.04)
Answer: dig

EXAMPLE 6
Context: A couple of trucks parked next to each other.
Question: What kind of company's do these trucks work for?
Candidates: construction (0.96), transport (0.57), haul (0.11), transportation (0.08), tow (0.04)
Answer: construction

Figure 1. **Answer candidates and answer-aware examples.** We show some typical samples of the testing inputs and their in-context examples. The predicted answers of Prophet have a high probability to appear in the answer candidates and answer-aware examples, showing the effectiveness of answer heuristics in enhancing GPT-3's ability to predict the correct answer. Although the samples in the last column are not answered correctly, their predictions are still reasonable under human evaluation.

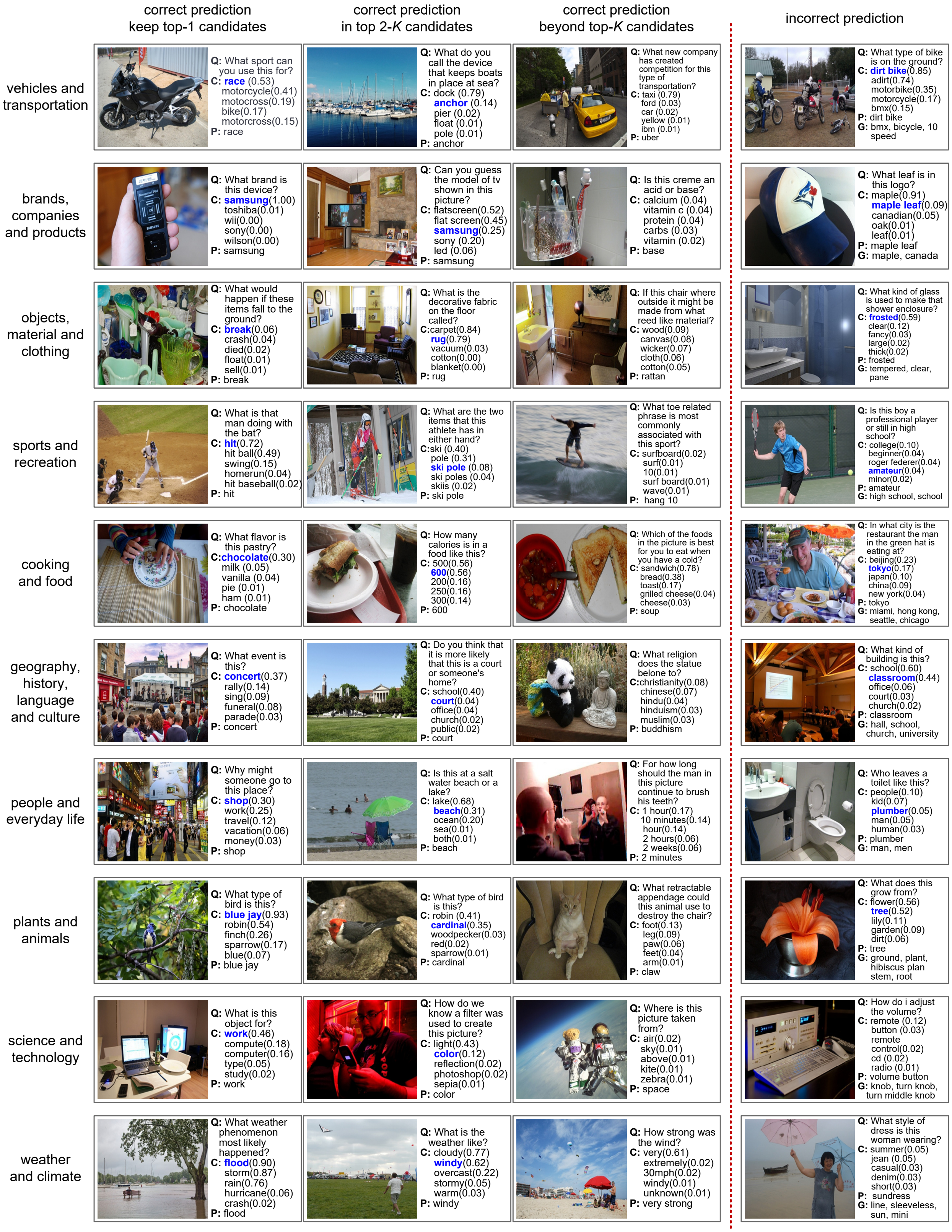

correct prediction keep top-1 candidates
correct prediction in top 2-K candidates
correct prediction beyond top-K candidates
incorrect prediction
vehicles and transportation
Q: What sport can you use this for?
C: race (0.53) motorcycle(0.41) motocross(0.19) bike(0.17) motorcross(0.15)
P: race
Q: What do you call the device that keeps boats in place at sea?
C: dock (0.79) anchor (0.14) pier (0.02) float (0.01) pole (0.01)
P: anchor
Q: What new company has created competition for this type of transportation?
C: taxi (0.79) ford (0.03) car (0.02) yellow (0.01) ibm (0.01)
P: uber
Q: What type of bike is on the ground?
C: dirt bike(0.85) adirt(0.74) motorbike(0.35) motorcycle(0.17) bmx(0.15)
P: dirt bike
G: bmx, bicycle, 10 speed
brands, companies and products
Q: What brand is this device?
C: samsung(1.00) toshiba(0.01) wii(0.00) sony(0.00) wilson(0.00)
P: samsung
Q: Can you guess the model of tv shown in this picture?
C: flatscreen(0.52) flat screen(0.45) samsung(0.25) sony (0.20) led (0.06)
P: samsung
Q: Is this creme an acid or base?
C: calcium (0.04) vitamin c (0.04) protein (0.04) carbs (0.03) vitamin (0.02)
P: base
Q: What leaf is in this logo?
C: maple(0.91) maple leaf(0.09) canadian(0.05) oak(0.01) leaf(0.01)
P: maple leaf
G: maple, canada
objects, material and clothing
Q: What would happen if these items fall to the ground?
C: break(0.06) crash(0.04) died(0.02) float(0.01) sell(0.01)
P: break
Q: What is the decorative fabric on the floor called?
C:carpet(0.84) rug(0.79) vacuum(0.03) cotton(0.00) blanket(0.00)
P: rug
Q: If this chair where outside it might be made from what reed like material?
C: wood(0.09) canvas(0.08) wicker(0.07) cloth(0.06) cotton(0.05)
P: rattan
Q: What kind of glass is used to make that shower enclosure?
C: frosted(0.59) clear(0.12) fancy(0.03) large(0.02) thick(0.02)
P: frosted
G: tempered, clear, pane
sports and recreation
Q: What is that man doing with the bat?
C: hit(0.72) hit ball(0.49) swing(0.15) homerun(0.04) hit baseball(0.02)
P: hit
Q: What are the two items that this athlete has in either hand?
C:ski (0.40) pole (0.31) ski pole (0.08) ski poles (0.04) skiis (0.02)
P: ski pole
Q: What toe related phrase is most commonly associated with this sport?
C: surfboard(0.02) surf(0.01) 10(0.01) surf board(0.01) wave(0.01)
P: hang 10
Q: Is this boy a professional player or still in high school?
C: college(0.10) beginner(0.04) roger federer(0.04) amateur(0.04) minor(0.02)
P: amateur
G: high school, school
cooking and food
Q: What flavor is this pastry?
C:chocolate(0.30) milk (0.05) vanilla (0.04) pie (0.01) ham (0.01)
P: chocolate
Q: How many calories is in a food like this?
C: 500(0.56) 600(0.56) 200(0.16) 250(0.16) 300(0.14)
P: 600
Q: Which of the foods in the picture is best for you to eat when you have a cold?
C: sandwich(0.78) bread(0.38) toast(0.17) grilled cheese(0.04) cheese(0.03)
P: soup
Q: In what city is the restaurant the man in the green hat is eating at?
C: beijing(0.23) tokyo(0.17) japan(0.10) china(0.09) new york(0.04)
P: tokyo
G: miami, hong kong, seattle, chicago
geography, history, language and culture
Q: What event is this?
C: concert(0.37) rally(0.14) sing(0.09) funeral(0.08) parade(0.03)
P: concert
Q: Do you think that it is more likely that this is a court or someone's home?
C: school(0.40) court(0.04) office(0.04) church(0.02) public(0.02)
P: court
Q: What religion does the statue belone to?
C:christianity(0.08) chinese(0.07) hindu(0.04) hinduism(0.03) muslim(0.03)
P: buddhism
Q: What kind of building is this?
C: school(0.60) classroom(0.44) office(0.06) court(0.03) church(0.02)
P: classroom
G: hall, school, church, university
people and everyday life
Q: Why might someone go to this place?
C: shop(0.30) work(0.25) travel(0.12) vacation(0.06) money(0.03)
P: shop
Q: Is this at a salt water beach or a lake?
C: lake(0.68) beach(0.31) ocean(0.20) sea(0.01) both(0.01)
P: beach
Q: For how long should the man in this picture continue to brush his teeth?
C: 1 hour(0.17) 10 minutes(0.14) hour(0.14) 2 hours(0.06) 2 weeks(0.06)
P: 2 minutes
Q: Who leaves a toilet like this?
C: people(0.10) kid(0.07) plumber(0.05) man(0.05) human(0.03)
P: plumber
G: man, men
plants and animals
Q: What type of bird is this?
C: blue jay(0.93) robin(0.54) finch(0.26) sparrow(0.17) blue(0.07)
P: blue jay
Q: What type of bird is this?
C: robin (0.41) cardinal(0.35) woodpecker(0.03) red(0.02) sparrow(0.01)
P: cardinal
Q: What retractable appendage could this animal use to destroy the chair?
C: foot(0.13) leg(0.09) paw(0.06) feet(0.04) arm(0.01)
P: claw
Q: What does this grow from?
C: flower(0.56) tree(0.52) lily(0.11) garden(0.09) dirt(0.06)
P: tree
G: ground, plant, hibiscus plan stem, root
science and technology
Q: What is this object for?
C: work(0.46) compute(0.18) computer(0.16) type(0.05) study(0.02)
P: work
Q: How do we know a filter was used to create this picture?
C: light(0.43) color(0.12) reflection(0.02) photoshop(0.02) sepia(0.01)
P: color
Q: Where is this picture taken from?
C: air(0.02) sky(0.01) above(0.01) kite(0.01) zebra(0.01)
P: space
Q: How do i adjust the volume?
C: remote (0.12) button (0.03) remote control(0.02) cd (0.02) radio (0.01)
P: volume button
G: knob, turn knob, turn middle knob
weather and climate
Q: What weather phenomenon most likely happened?
C: flood(0.90) storm(0.87) rain(0.76) hurricane(0.06) crash(0.02)
P: flood
Q: What is the weather like?
C: cloudy(0.77) windy(0.62) overcast(0.22) stormy(0.05) warm(0.03)
P: windy
Q: How strong was the wind?
C: very(0.61) extremely(0.02) 30mph(0.02) windy(0.01) unknown(0.01)
P: very strong
Q: What style of dress is this woman wearing?
C: summer(0.05) jean (0.05) casual(0.03) denim(0.03) short(0.03)
P: sundress
G: line, sleeveless, sun, mini

Figure 2. **Different categories and prediction behaviors.** Each row contains four testing samples from a specific knowledge category. The first to the third columns correspond to the correctly answered samples of different prediction behaviors (*i.e.*, keep top-1, in top 2-K, and beyond top-K). The last column contains failure samples.