Capturing Wonder

Bridging Imagery and Text in Children's Books

Crystal Li, Laura Li, Kevin Sheng, Sean Huang

Agenda

01
Problem Description

02
Approach

03
Results

04
Lessons Learned

Problem Description

Image Captioning

Image captioning blends **computer vision** with **natural language processing** to create textual descriptions for images.

It bridges the **visual-textual gap** and makes reading more inclusive. It holds particular promise for young readers with visual impairments or learning disabilities.

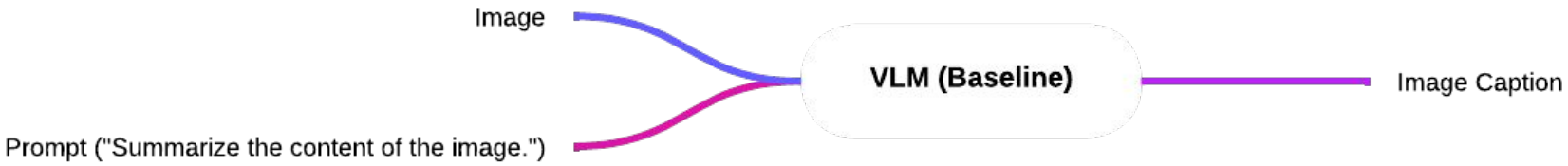
It can also **improve classification systems** across libraries, bookstores, and online platforms, providing clearer insights into book content.

Our Approach

- Pre-trained model (VLM)
- GPT-4-Vision-Preview
- GPT-4-Vision-Preview and fine-tuned GPT-3.5-turbo

Pre-trained Model – Vision Language Model (VLM)

- Multi-modal model that contains:
  - CNN to obtain image embedding
  - Llama-1.2b for NLP task
    - Prompt: “Summarize the content of the image”

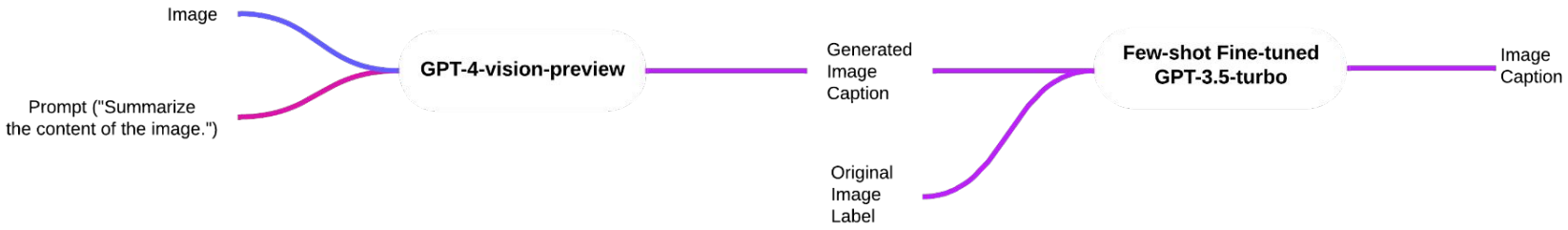


OpenAI – GPT-4-Vision-Preview

- Image converted to base64 format and fed to gpt-4 along with prompt
  - Prompt: “Summarize the content of the image”
- Received image caption in response

OpenAI –
GPT-4 and Fine-tuned GPT-3.5

- Added a layer of GPT-3.5-turbo to the pipeline
  - Idea is to make the output of the GPT-4 model sound like the original image label
  - Fine-tuning train dataset consists of generated caption from GPT-4 and the original image label provided by the dataset



Method of Evaluation

BERT Embeddings

Get BERT embeddings for both the actual label and generated caption

Cosine-Similarity Matrix

Cosine-similarity for each component in the embedding

Average Similarity Score

Take average of cosine-similarity matrix minus the diagonal self-similarity

Results

Cos Similarity	VLM Baseline	GPT-4-Vision-Preview	Few-shot fine-tuned GPT
Min Score	0.6113	0.4761	0.4305
Max Score	0.8802	0.8960	0.9434
Average Score	0.7747	0.8067	0.7987

- GPT-4 shows a **broader score range**, but occasionally lower alignment with original captions.
- Higher max and average scores with GPT-4 suggest greater capacity for generating closely aligned captions.
- **Better average performance** of GPT-4 likely due to its larger training size and more sophisticated architecture.

Fine-tuning Results

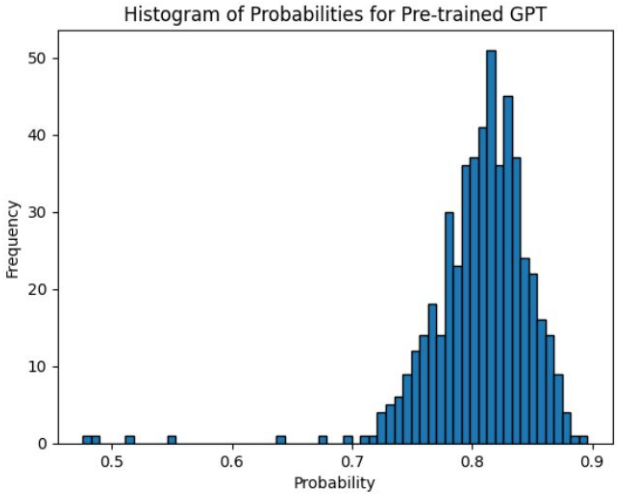


Figure 3: Pre-trained GPT

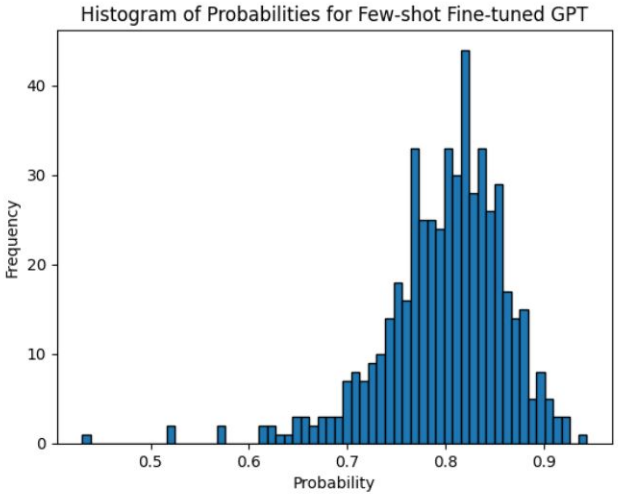


Figure 4: Fine-tuned GPT

- Fine-tuned model shows **more dispersed score distribution**; mean score remains similar.
- Fine-tuning for text shortening leads to **more extreme** output quality.
- Advanced modeling results in an unexpected magnification effect on score dispersion.

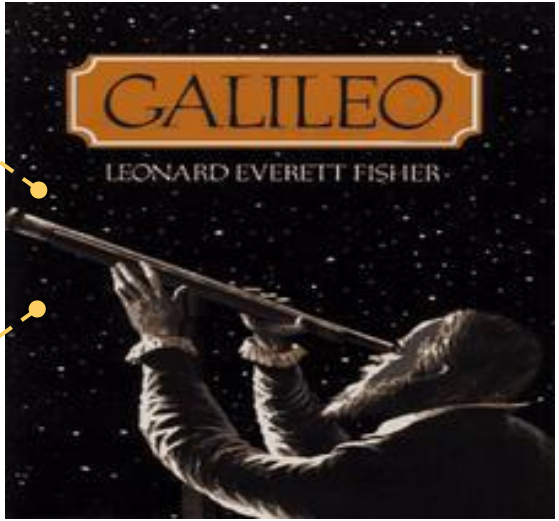
Generated Caption
Example 1

Original Label:
Drawing of a man with a telescope looking at stars with the title 'GALILEO'

VLM Baseline Model:
A man sits on a bed, holding a gun and aiming it at a target. He seems to be in a video game, possibly a first-person shooter, as he is dressed in a suit and holding a gun. The setting is a dark room, with the man as the primary subject of the image

1

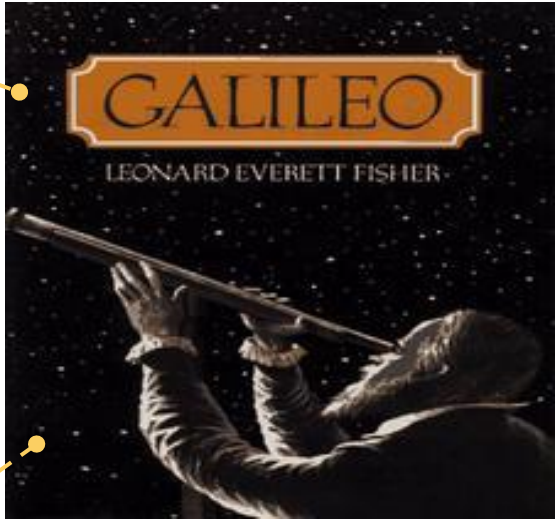
2



Generated Caption
Example 1

GPT-4-Vision-Preview model:
The image shows a person looking up at the stars in the night sky through a long telescope. The background is dark with lots of little white dots that look like stars. Above the person, there's a big word "GALILEO" in gold letters, which is the name of a very famous person who studied the stars a long time ago. It looks like the cover of a book that tells the story of Galileo.

3

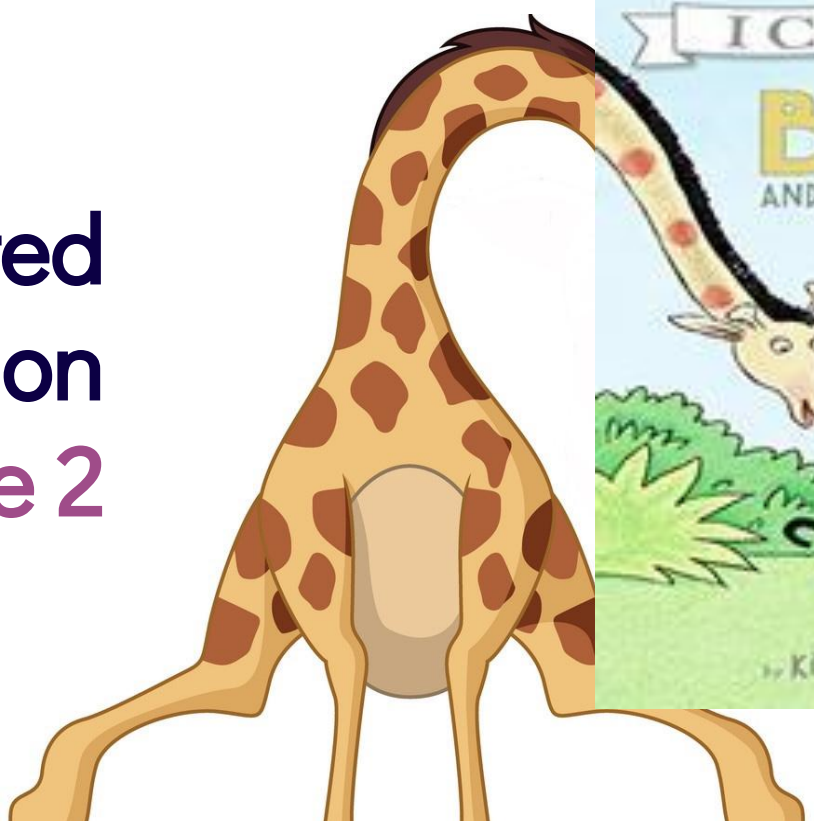


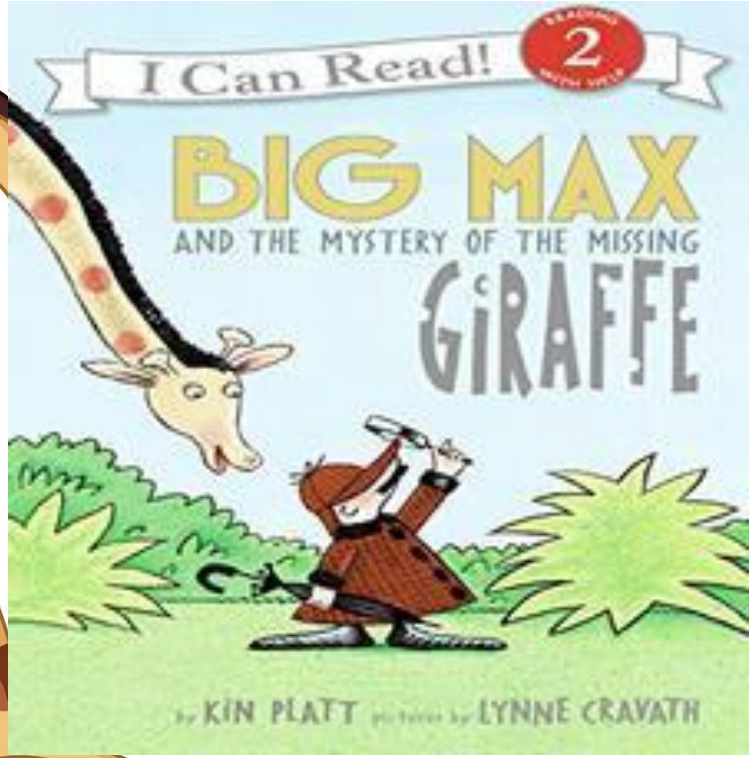
4

Few-shot Fine-tuned Model:
Whispers of the Cosmos: The Stargazing Tales of Galileo.

Generated Caption

Example 2





Original Label:
Drawing of a detective in brown clothes holding a magnifying glass in a jungle with a giraffe looming from the left side behind him and the caption 'Big Max and the mystery of the missing giraffe.

VLM Baseline Model:
A cartoon depicts a giraffe and a man sitting together in a grassy area. The giraffe is on the left, and the man is on the right. The scene is set in a park, with a potted plant in the background. The giraffe holds a book, possibly a children's book, in its mouth. The man appears to be reading the book, creating a playful and lighthearted atmosphere.

GPT-4-Vision-Preview model:
This is a picture of a book cover. It shows a drawing of a detective who looks like he's trying to solve a mystery. The detective, Big Max, has a magnifying glass and seems to be searching for clues. There's a tall giraffe peeping into the picture, which might be part of the mystery he's trying to solve. It looks like a fun story for kids to read about a detective on an adventure to find a missing giraffe.

Few-shot Fine-tuned Model:
Illustration of a giraffe and a detective named Big Max with the title 'Big Max and the Missing Giraffe.'

Lessons Learned

Domain-Specific Training

Fine-tuning on domain-specific data enhances model's ability to capture stylistic elements suitable for children's literature.

Interdisciplinary Collaboration

Ensures captions are not only accurate but also meet cognitive and developmental needs of young readers.

Comprehensive Metrics

Traditional metrics may not capture subjective quality (creativity, coherence, suitability, etc.)

Ethical Considerations

Awareness of cultural sensitivities in AI-generated content. Responsible AI involves monitoring & evaluation to ensure inclusivity & positive impact.

Thank You!