5th International Symposium on AIE

Instruction-Following Evaluation for Large Vision-Language Models (LVLMs)

Daiki Shiono, 2nd year of master program, Graduate School of Information Sciences

Abstract

- Creation of fine-tuning datasets **with instructions on output format**
- For the first time, **quantitatively** demonstrated a decrease in LVLM's ability to follow instructions after fine-tuning
- **The presence or absence of instructions regarding the output format at the time of fine-tuning** is likely to have a significant impact on the LVLM's ability to follow instructions

Proposed Method

Create (Visual) Instruction Tuning Datasets

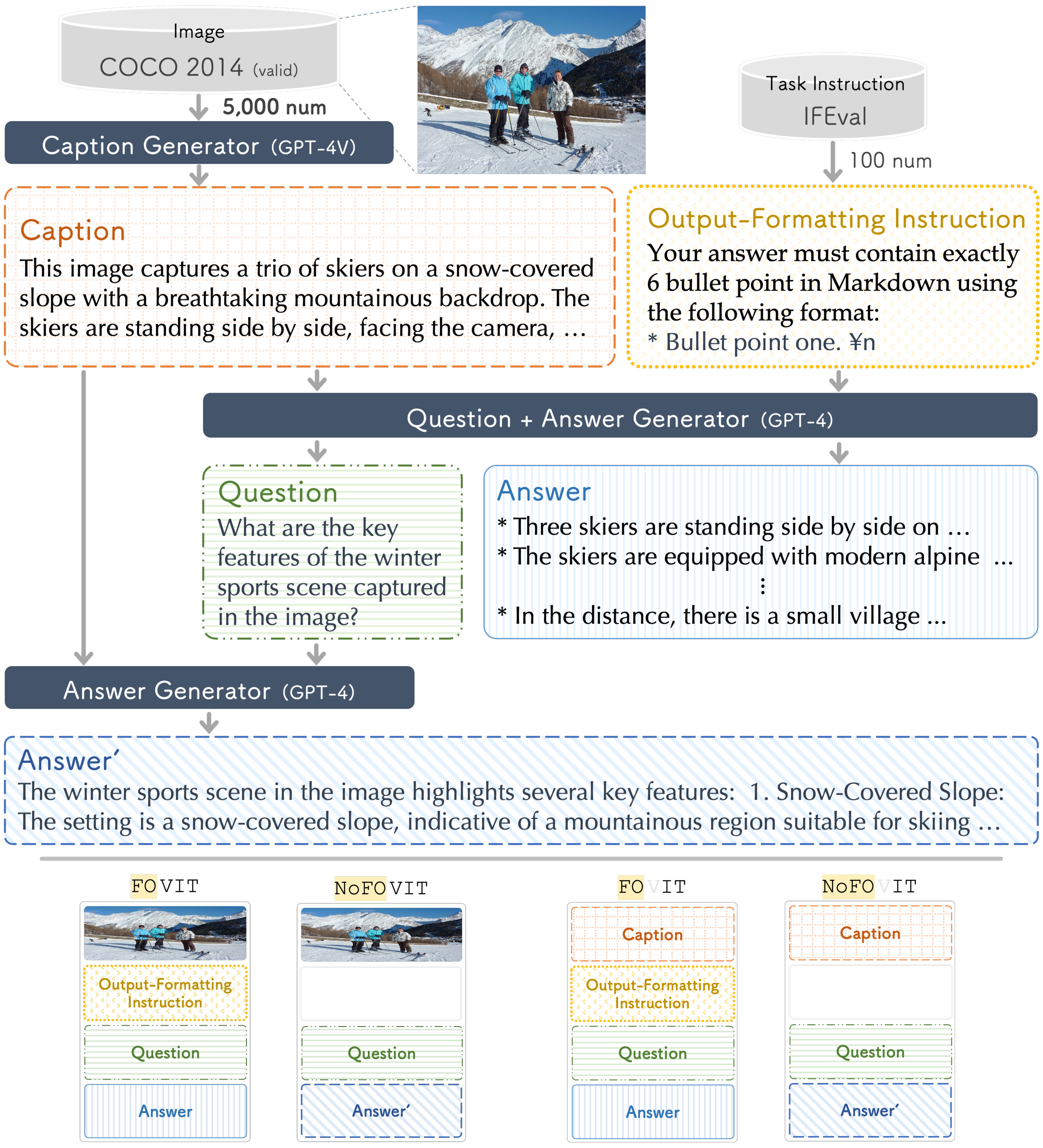


Image
COCO 2014 (valid)
5,000 num
Caption Generator (GPT-4V)

Task Instruction
IFEval
100 num

Caption
This image captures a trio of skiers on a snow-covered slope with a breathtaking mountainous backdrop. The skiers are standing side by side, facing the camera, …

Output-Formatting Instruction
Your answer must contain exactly 6 bullet point in Markdown using the following format:
* Bullet point one. ¥n

Question + Answer Generator (GPT-4)

Question
What are the key features of the winter sports scene captured in the image?

Answer
* Three skiers are standing side by side on …
* The skiers are equipped with modern alpine …
⋮
* In the distance, there is a small village …

Answer Generator (GPT-4)

Answer'
The winter sports scene in the image highlights several key features: 1. Snow-Covered Slope: The setting is a snow-covered slope, indicative of a mountainous region suitable for skiing …

Background

- Cases was observed where the LVLM doesn't follow task instructions without showing the instruction-following ability that LLM had before it was incorporated. [Fu+, '23]
- We observe that existing Visual Instruction Tuning datasets often do not include instructions regarding output format.

Create Datasets for Evaluation of Instruction-Following Ability

# Examples of evaluation datasets

If a movie review is **positive**,
you need to output "{label_0}".
If a movie review is **negative**,
you need to output "{label_1}".

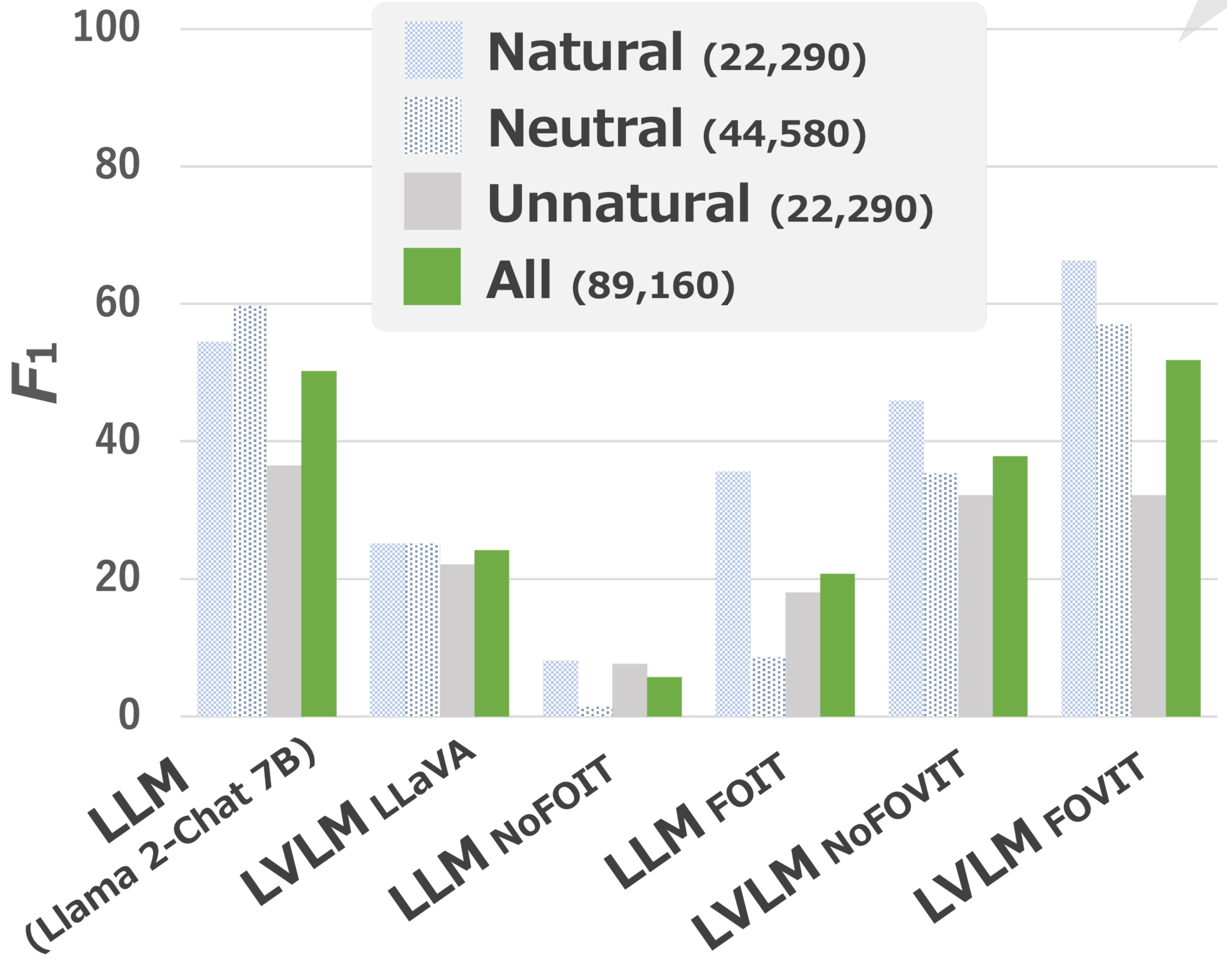
Movie review: lovely and poignant.
Answer:

Label System by Contextual Consistency		If positive.. label_0	If negative.. label_1
Natural	high	positive	negative
Neutral	↕	foo	bar
Unnatural	low	negative	positive

- Following Li et al. [Li+, '23], we performed **verbalizer manipulation** on each of the nine binary classification datasets (SST-2, FP, EMOTION, SNLI, SICK, RTE, QQP, MRPC, SUBJ) to construct evaluation datasets.
- Define 3 label systems according to the consistency between the semantic representation of the label and the contextual knowledge at the time of fine-tuning.

Experiment

LVLM Components :
・Llama 2-Chat 7B ・CLIP ViT-Large/14 ・1 Linear Layer



※ "All" indicates the macro average of F1 for "Natural", "Neutral", and "Unnatural".

✔ Quantitatively confirmed the decline in LVLM's instruction-following ability

- In "Unnatural", all LVLMs were below their base LLM (Llama 2-Chat 7B).

✔ Influenced by the presence or absence of instructions regarding the output format in the fine-tuning datasets

- LVLM FOVIT is higher F1 score than LVLM NoFOVIT.
- LLM FOIT is higher F1 score than LLM NoFOIT.
- Suggests that **explicitly giving the instructions on output format** can suppress the decline in the instruction-following ability that the base LLM possesses, regardless of modalities.