Supplementary Material for mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model

Anonymous Authors
Submission ID: 3283

ACM Reference Format:
Anonymous Authors, Submission ID: 3283. 2018. Supplementary Material for mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 M-PAPER

1.1 Outline Construction

In the scenario of assisted essay writing, the 'outline' given by users could be multiple content-related key points or a highly concise summary, such as 'the overall architecture of our model'. To simulate such diverse inputs, in M-Paper, we construct two types of outlines by designing different prompts and in-context demonstrations for GPT-3.5, as shown in Table 2 and Table 3.

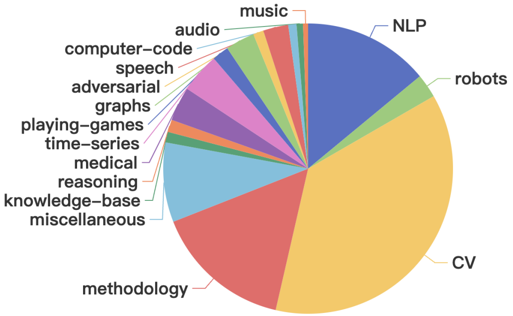


Figure 1: The category distribution of 48,688 academic papers.

1.2 Statistic

The detailed category distribution of papers in M-Paper is shown in Fig. 1.

1.3 Task Instruction

As shown in Table 4, for each task, we design diverse instructions to enhance the general instruction-following ability of the model.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 GPT-BASED METRIC

For evaluating the overall semantic similarity of a predicted diagram analysis and ground-truth one, we design a GPT-based metric, namely $F1^{gpt}$. We first prompt GPT to extract key points of prediction and ground truth. Then, for each pair of predicted key point and ground-truth one, we further prompt GPT to judge whether it matches or not. Finally, based on GPT's judgments, we calculate the precision, recall, and F1 score ($F1^{gpt}$). The prompts used in these two steps are shown in Table 5. In particular, during the keypoint extraction process, we prompt GPT to simultaneously process both the prediction and the ground truth to better capture their similarities and differences.

3 EXPERIMENTS

3.1 Influence of Table Format

For developing a copilot capable of analyzing different formats of diagrams during paper-writing, M-Paper evaluates table understanding in both image and Latex formats. As shown in Table 1, for writing a caption to summarize the table content, understanding Latex is much easier than understanding the image because all data is well-organized in text. However, the Latex format doesn't bring significant improvement for *Multimodal Diagram Anaylysis* and even a decrease in the CIDEr score. This is because when provided latex code of a table, the model tends to describe more rare prop nouns or numbers in the table, which may not be necessary for the discussion and don't appear in the ground-truth analysis. This shows that generating diagram analysis is more challenging at correlating [$Context$], [$Outline$], and [$Diagrams$], rather than mainly understanding the diagram content.

Table 1: The Multimodal Diagram Captioning and Analysis performance on .

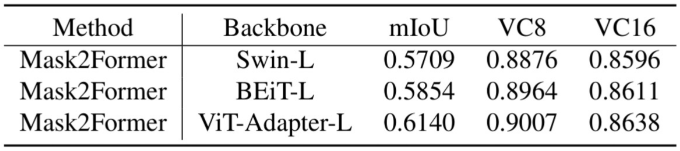
Table	Captioning			Analysis					
Format	R	M	C	B4	R	M	C	$F1^{gpt}$	C^{gpt}
Image	22.51	9.60	51.77	12.25	30.52	17.83	25.64	0.45	14.50
Latex	26.69	10.54	80.03	12.03	30.38	18.11	21.56	0.47	10.51

3.2 More Qualitative Results

Fig. 2 and Fig. 3 show more qualitative results of Multimodal Diagram Captioning, including multiple sub-figures and tables in the image or latex code format.

During outline construction, the average length of the [$Outline$] is around 36.43% of the target diagram analysis to avoid leaking too much information about diagrams. Although we perform such data preprocess to enforce the model learning to describe more diagram details during training, PaperOwl sometimes still makes little revisions based on the outline and doesn't provide informative analysis about the diagram, as presented in Fig. 4. Therefore, how to instruct the model to follow the outline and encourage it to give more details by understanding diagrams is also a challenge for Multimodal Diagram Analysis.

[***Diagrams***]



Method	Backbone	mIoU	VC8	VC16
Mask2Former	Swin-L	0.5709	0.8876	0.8596
Mask2Former	BEiT-L	0.5854	0.8964	0.8611
Mask2Former	ViT-Adapter-L	0.6140	0.9007	0.8638

Table 1

[***Context***]: With the proposed method, Our solution not only performs well in mIoU, but also outperforms others significantly in video frame continuity and consistency on the PVUW2023 challenge.In this paper, ...Finally, we obtain the 2nd place in the final test set, as shown in Table 4.

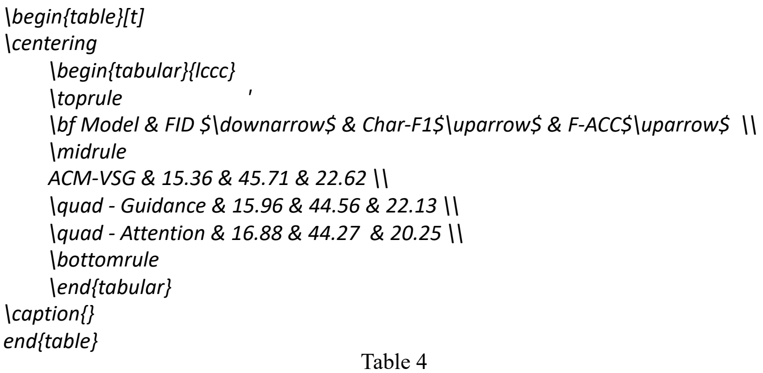
[***Insturction***]: 'Give a short caption that accurately conveys the main idea of this figure.

GT: Experiments of different backbones on PVUW2023 challenge test part 1

PaperOwl ***:*** The comparison of different backbones on the PVUW2023 validation set.

(a)

[***Diagrams***]



```
\begin{table}[t]
\centering
      \begin{tabular}{lccc}
      \toprule                          '
      \bf Model & FID $\downarrow$ & Char-F1$\uparrow$ & F-ACC$\uparrow$  \\
      \midrule
      ACM-VSG & 15.36 & 45.71 & 22.62 \\
      \quad - Guidance & 15.96 & 44.56 & 22.13 \\
      \quad - Attention & 16.88 & 44.27  & 20.25 \\
      \bottomrule
      \end{tabular}
\caption{}
end{table}
```

Table 4

[***Context***]: Table 2 shows the results for story continuation task. As we can see, our model can achieve the best results on both datasets, 15.36 and 18.41 FID for PororoSV and FlintstonesSV, respectively. And our model can greatly preserve characters to improve the consistency of the story., …… includes a recurrent text encoder, an image generation module, and two discriminators - image and story discriminator.

[***Insturction***]: I need a succinct caption for this diagram.

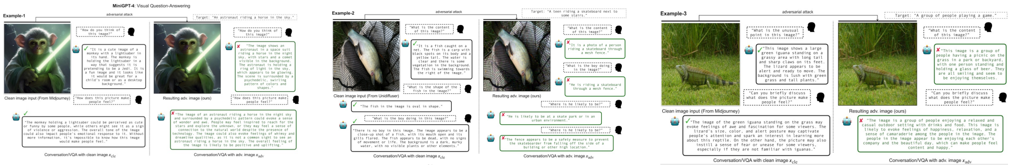
GT: Ablation study results for story continuation task on PororoSV.

PaperOwl ***:*** Ablation study on PororoSV dataset for story continuation task.

(b)

Figure 2: Qualitative results of PaperOwl for captioning tables in the image format (a) and Latex format (b). Wrong descriptions are marked as red.

[*Diagrams*]



MiniGPT-4: Visual Question-Answering

Example-1

adversarial attack

Target: "An astronaut riding a horse in the sky."

"How do you think of this image?"

✓ "It is a cute image of a monkey with a lightsaber in its hand. The monkey is holding the lightsaber in a way that suggests it is pretending to be a Jedi. It is a fun image and it looks like it would be great for a child's room or as a desktop background."

Clean image input (From Midjourney)

"How does this picture make people feel?"

✓ "The monkey holding a lightsaber could be perceived as cute or funny by some people, while others might see it as a sign of violence or aggression. The overall tone of the image could also impact people's emotional response to it. Without more information, it's impossible to know how this image would make people feel."

Conversation/VQA with clean image x_{cle}

"How do you think of this image?"

✗ "The image shows an astronaut in a space suit riding a horse in the night sky, with stars and a comet visible in the background. The astronaut is holding a ring of light in the sky, which appears to be glowing. The scene is surrounded by a psychedelic, swirling pattern of colors and shapes."

Resulting adv. image (ours)

"How does this picture make people feel?"

✗ "The image of an astronaut riding a horse in the night sky and surrounded by a psychedelic pattern could evoke a sense of wonder and awe. People may feel inspired to reach for the stars and explore the unknown, or they may feel a sense of connection to the natural world despite the presence of technology. The image could also evoke feelings of whimsy and dreamlike qualities, as it is not a common sight to see an astronaut riding a horse in the sky. The overall feeling of the image is likely to be positive and uplifting."

Conversation/VQA with adv. image x_{adv}

Example-2

adversarial attack

Target: "A teen riding a skateboard next to some stairs."

"What is the content of this image?"

✓ "It is a fish caught on a net. The fish is a carp with black spots on its body and a yellow tail. The water is clear and there is some vegetation in the background. The fish is swimming towards the right of the image."

"What is the shape of the fish in the image?"

Clean image input (From Unidiffuser)

✓ "The fish in the image is oval in shape."

"What is the boy doing in this image?"

✓ "There is no boy in this image. The image appears to be a close-up shot of a fish, with its mouth open and its gills flared. The fish appears to be dead, with no signs of movement or life. The background is a dark, murky water, with no visible plants or other elements."

Conversation/VQA with clean image x_{cle}

"What is the content of this image?"

✗ "It is a photo of a person riding a skateboard through a mesh fence."

"What is the boy doing in the image?"

✗ "He is riding a skateboard through a mesh fence."

Resulting adv. image (ours)

"Where is he likely to be?"

✗ "He is likely to be at a skate park or in an urban environment."

"Where is he likely to be?"

✗ "The fence appears to be a safety measure to prevent the skateboarder from falling off the side of a building or other high location."

Conversation/VQA with adv. image x_{adv}

Example-3

adversarial attack

Target: "A group of people playing a game."

"What is the unusual point in this image?"

✓ "This image shows a large green iguana standing on a grassy area with long tail and sharp claws on its feet. The lizard appears to be alert and ready to move. The background is lush with green grass and tall plants."

"Can you briefly discuss what does the picture make people feel?"

Clean image input (From Midjourney)

✓ "The image of the green iguana standing on the grass may evoke feelings of awe and fascination for some viewers. The lizard's size, color, and alert posture may captivate people's attention and spark an interest in learning more about this reptile. On the other hand, the picture may also instill a sense of fear or unease for some viewers, especially if they are not familiar with iguanas."

Conversation/VQA with clean image x_{cle}

"What is the content of this image?"

✗ "This image is a group of people having a picnic on the grass in a park or backyard, with one person standing and holding a glass of beer. They are all smiling and seem to be enjoying themselves."

"Can you briefly discuss what does the picture make people feel?"

Resulting adv. image (ours)

✗ "The image is a group of people enjoying a relaxed and casual outdoor setting with drinks and food. This image is likely to evoke feelings of happiness, relaxation, and a sense of camaraderie among the people in the image. The people in the image appear to be enjoying each other's company and the beautiful day, which can make people feel content and happy."

Conversation/VQA with adv. image x_{adv}

figure 11

[***Context***]: Unidiffuser [4] models the joint generation across multiple modalities, such as text-to-image or image-to-text generation. In Figure 10, we show additional results for the joint generation task implemented by Unidiffuser. As can be seen, our crafted adversarial examples elicit the targeted response in various generation paradigms.... In this section, we present additional results, visualization, and analyses to supplement the findings in our main paper.

[***Insturction***]: Describe these images concisely.

GT: Additional results of VQA task implemented by MiniGPT-4.

PaperOwl ***:*** Additional results for the question-answering task.

(a)

[*Diagrams*]

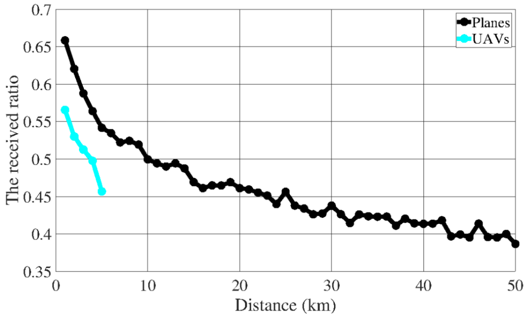


figure 6

[***Context***]: As demonstrated in Fig. 5 , each plane is added with the distance factor, and the packets propagation is added with the channel model. After adding the distance factor, apart from the loss of packets caused by collision, error codes caused by propagation loss also packets discarded. ...If the requirement of updating interval is 3s, i.e., more than 6 consecutive packets loss are regarded as the position is not updated. The position updating probability within 3s is 97.8 $\\%$.

[***Insturction***]: I need a succinct caption for this figure..

GT: The received ratio on planes and UAVs at different distance.

PaperOwl ***:*** The received ratio of different planes and UAVs.

(b)

Figure 3: Qualitative results of PaperOwl for captioning figures with multiple sub-images (a) and only 1 image (b). Wrong descriptions are marked as red.

[*Diagrams*]

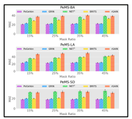
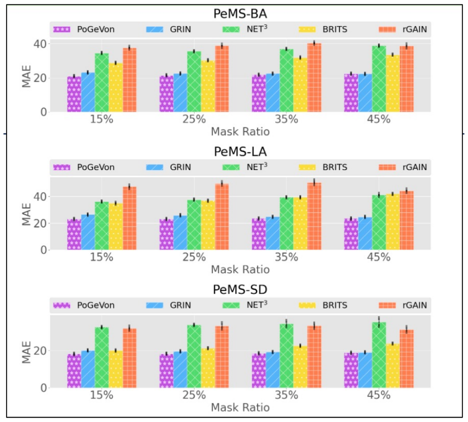


figure 8: figure 4: Sensitivity analysis for time series imputation with different masking rates on the traffic dataset. Lower is better. Best viewed in color.



[***Context***]**:** We conduct sensitivity analysis to study the effect brought by increasing the masking rates. We consider the following mask rates: $15\\%$, $25\\%$, $35\\%$, $45\\%$ …… . This indicates that both VGAE and VGRNN may not be able to handle the link prediction task over weighted dynamic graphs very well.

[***Outline***]**:** The proposed PoGeVon method consistently outperforms all baselines in various settings, demonstrating its effectiveness and robustness.

[***Insturction***]**:** With reference to the previous content, the summary and these illustrations, provide a comprehensive and articulate paragraph analysis.

GT: The results are shown in Figure 4 , in which the error bar demonstrates the standard deviation of MAE over 5 runs with different random seeds. The proposed PoGeVon consistently outperforms all the baselines in these settings which further demonstrates the effectiveness and robustness of our method.

PaperOwl ***:*** The results are shown in Figure 4 . We can see that the proposed PoGeVon consistently outperforms all baselines in all the settings. This demonstrates the effectiveness and robustness of our proposed method.

Figure 4: A failure case of PaperOwl for Multimodal Diagram Analysis. Key messages coming from diagrams are marked as green.

Table 2: Prompts used for generating a highly concise 'outline'.

Please provide the main point of the following paragraph which is from a scientific paper. The main point is the central issue in the paragraph and the format like some items in the outline, and it should be as concise and brief as possible!!!!

Due to the paragraph being from a scientific paper, it can be like: the background of some tasks, or the challenge of previous methods, our methods involve A and B modules, etc for the paragraph from the Introduction section; or experiments results on some datasets for the paragraph from Experiments section, or the pipeline of feature extractor, or the detailed design of some network for the paragraph from Method section.

Please provide a highly abstract writing purpose for this paragraph like an outline, rather than simply summarizing the content of the paragraph.

And please generate the main point with less than 20 words! less than 20 words! less than 20 words!!!

There are some examples of "Paragraph" and "Main Points" pairs. The examples are split by "##############################":

##############################
Paragraph:
\noindent \textbf{Low Reference Dependency} The Kendall and Spearman correlations between automatic metrics and human judgments with the different numbers of references are shown in Fig.\ref{fig:changing_reference_number}. Our EMScore without any references can achieve competitive results, compared with reference-based metrics which need at least 4 or 5 references, such as BLEU_1 and Improved_BERTScore. Besides, our EMScore_ref with only one reference can achieve comparable results with reference-based metrics, which need at least 8 or 9 references, such as CIDEr and BERTScore. The results show that our metric has lower reference dependency, which benefits from the introduction of video content in evaluation.

Main Points:
Our metric has a lower reference dependency.
##############################
Paragraph:
Fig.\ref{fig:fine_grained_matching} visualizes how fine-grained EMScore matches the most similar visual elements to the tokens (as the calculation of precision). For the first example, "bubbles" occurs in the 106th frame, "another boy" occurs in the 160th and 187th frames, and compared with other frames, "face paint" appears in a larger proportion in the 4th and 6th frames. For the second example, the visual concept "boy" appears as the main visual element in the 53rd frame, so the token 'boy' matches this frame instead of 84th$\sim$298th frames where multiple visual elements appear. Compared with coarse-grained embedding matching, our fine-grained one can take into account the characteristics of the video, and provide more interpretability for EMScore.

Main Points:
The visualization results of fine-grained EMScore.
##############################

Paragraph: [$Paragraph$]
Main Points: [$Main\ Points$]

Table 3: Prompts used for generating an 'outline' in the form of multiple key points.

```
Please use one or several concise sentences to summarize the main points of the following paragraph which is from a scientific paper.
And please note that:
(1) Each sentence should strive to express one main point as succinctly as possible.
(2) Please summarize the most critical points, preferably no more than 3. And one main point is enough for some short paragraphs!!!
(3) If there are multiple main points, use "1. 2. 3." to list them and use "\n" to split them.

There are some wrong formats with prefix like this: "The article introduces xxx".
"The authors conduct experiments xxx".
"They introduce xx".
"xxx proposed by the author".
Please directly generate the key points of the paragraph, and don't use the prefix like above.

There are some examples of "Paragraph" and "Main Points" pairs. The examples are split by "##############################":

##############################
Paragraph:
Video Captioning\cite{DBLP:journals/tcsv/DengLZWZH22} aims to generate a text describing the visual content of a given video. Driven by the neural encoder-decoder paradigm, research in video captioning has made significant progress \cite{DBLP:conf/iccv/VenugopalanRDMD15, DBLP:conf/cvpr/ZhangSY0WHZ20}. To make further advances in video captioning, it is essential to accurately evaluate generated captions. The most ideal metric is human evaluation while carrying human judgments is time-consuming and labor-intensive. Thus, various automatic metrics are applied for video caption evaluation.

Main Points:
Accurately evaluating the generated descriptions is necessary, and due to the time-consuming and labor-intensive nature of human judgments, automatic evaluation metrics are widely used.

##############################
Paragraph:
However, most of the widely applied video caption metrics like BLEU\cite{DBLP:conf/acl/PapineniRWZ02}, ROUGE\cite{lin-2004-rouge}, CIDEr\cite{7299087}, and BERTScore\cite{DBLP:conf/iclr/ZhangKWWA20} come from the other tasks, such as machine translation, text summarization and image captioning, which may neglect the special characteristic of video captioning and then limit the development of video captioning. Furthermore, these automatic metrics require human-labeled references — and thus they are called reference-based metrics — and such requirements cause three intrinsic drawbacks: (1) They can not be used when provided videos have no human-labeled references, which is not uncommon in this age that millions of reference-free videos are produced online every day. (2) They may over-penalize the correct captions since references hardly describe all details of videos due to the one-to-many nature\cite{DBLP:conf/acl/YiDH20} of captioning task, especially when the number of references is limited. Fig.\ref{fig:introductionexample} (a) shows one such example where a candidate caption correctly describes the "a rock" while reference-based metrics punish this word since references do not contain it. (3) As pointed by \cite{rohrbach-etal-2018-object}, these reference-based metrics may under-penalize the captions with "hallucinating" descriptions since these metrics only measure similarity to references, and the visual relevance cannot be fully captured. For example, as shown in Fig.\ref{fig:introductionexample} (b), due to the word "games" appearing in the references, some reference-metrics return higher scores for caption B than caption A, even though "different games" is a "hallucinating" phrase which is not related to the video.

Main Points:
1. Commonly used video caption metrics come from other tasks and may not fully capture the unique characteristics of video captioning.
2. The requirement of reference causes three intrinsic drawbacks: (1) Cannot be applied in real time. (2) Over-penalize the correct captions. (3) Under-penalize the captions with "hallucinating" descriptions.
##############################

Paragraph: [Paragraph]
Main Points: [Main Points]
```

Table 4: Instructuion used for Multimodal Diagram Captioning, Multimodal Diagram Analysis and Outline Recommendation. The $[object]$ is randomly chosen from $\{figures, images, photos, pictures, diagrams, illustrations\}$ or $\{figure, image, photo, picture, diagram, illustration\}$ depending on the number of diagrams is more than 1 or not.

Multimodal Diagram Captioning
Describe $[object]$ concisely.
Write a caption of $[object]$.
Provide a brief description of $[object]$.
Write a short caption for $[object]$.
come up with a concise caption that captures the essence of $[object]$.
Encapsulate the key information presented in $[object]$ in a brief statement.
I need a succinct caption for $[object]$.
Please provide a pithy summary of $[object]$ that effectively communicates its message.
Can you provide a snappy caption that perfectly encapsulates the message conveyed by $[object]$?
Please write a brief but compelling caption that grabs the reader's attention and draws them into $[object]$.
Give a short caption that accurately conveys the main idea of $[object]$.
Multimodal Diagram Anaysis
Based on the previous content and the outline, write a detailed and fluent paragraph analysis.
With reference to the preceding content and the given summary, compose a comprehensive and articulate paragraph analysis.
Considering the information provided earlier and following the provided outline, produce a detailed and fluent analysis in paragraph form.
Drawing from the preceding content and adhering to the outlined structure, write a thorough and coherent paragraph analysis.
Based on the previous content and guided by the summary, construct a detailed and fluid analysis in paragraph format.
Taking into account the preceding information and following the provided outline, generate a comprehensive and well-developed paragraph analysis.
Considering the content discussed earlier and following the provided outline, present a detailed and fluent analysis in paragraph form.
With reference to the previous content and the summary, provide a comprehensive and articulate paragraph analysis.
Based on the preceding discussion and in accordance with the outlined structure, compose a detailed and coherent paragraph analysis.
Considering the information presented earlier and adhering to the provided summary, formulate a thorough and seamless paragraph analysis.
Outline Recommendation
more than 1 diagrams
Based on the previous content and $[object]$, list some key points that should be covered in the next paragraph.
Considering the preceding text with $[object]$, the next paragraph needs to address these essential aspects.
Drawing from the preceding text and image information, what crucial points should be focused on in the ensuing paragraph?
Given the multimodal information provided earlier, write some key factors for the next paragraph.
With reference to the previous discussion and $[object]$, the next paragraph should discuss the following important elements.
In light of the preceding content with $[object]$, which significant points should be analyzed in the subsequent paragraph?
Based on the previous text and $[object]$, the next paragraph should delve into these core aspects.
Considering the text and vision information presented before, give some main factors that should be addressed in the ensuing paragraph.
Taking into account the preceding discussion and $[object]$, what primary points should be emphasized in the next paragraph?
Given the previous context with $[object]$, generate some key elements that should be discussed in the next paragraph should discuss.
no diagrams
Based on the previous content, list some key points that should be covered in the next paragraph.
Considering the preceding text, the next paragraph needs to address these essential aspects.
Drawing from the preceding information, what crucial points should be focused on in the ensuing paragraph?
Given the information provided earlier, write some key factors for the next paragraph.
With reference to the previous discussion, the next paragraph should discuss the following important elements.
In light of the preceding content, which significant points should be analyzed in the subsequent paragraph?
Based on the previous text, the next paragraph should delve into these core aspects.
Considering the information presented before, give some main factors that should be addressed in the ensuing paragraph.
Taking into account the preceding discussion, what primary points should be emphasized in the next paragraph?
Given the previous context, generate some key elements that should be discussed in the next paragraph should discuss.

Table 5: Prompts used for calculate $F1^{gpt}$. $[Prediction]$ and $[Ground\ Truth]$ are predicted analysis and ground-truth analysis, respectively. $[Predicted\ Point]$ and $[GT\ Point]$ is a pair of key points extracted from the $[Prediction]$ and $[Ground\ Truth]$, respectively,

Prompt GPT for Extracting Key Points
Please summarize the main points of the prediction and ground truth. And strictly with the format: 1. xxx. 2. xxx. ... Please ensure that the generated main points comprehensively condense the information of the original text (prediction or ground truth). The number of generated main points can be as many as possible, but no more than 10. If there are parts of the prediction and ground truth that are the same, reflect that in main points, such as some main points of them are the same, and other main points summarize the unique content of themselves. Please note that if there are any overlapping contents between the prediction and ground truth, the main points for these contents should remain consistent. However, for different content of them, please provide separate main points for each. The format is as follows: ####### Predicted text: xxx. Ground Truth text: xxx. The main points of the predicted text: 1. xx 2. xx ... The main points of the ground truth text: 1. xx 2. xx ... ####### Now, please generate the main points of the given prediction and ground truth, please strictly use the prompt 'The main points of the xxx' in the response. Predicted text: $[Prediction]$ Ground Truth text: $[Ground\ Truth]$
Prompt GPT for Judging Semantic Matching
Given a predicted text and a reference text, please judge whether the semantics of the predicted text can match the reference text. And use Yes or No to represent match or mismatch. The format is as follows: Predicted text: xxx. Reference text: xxx. Yes/No ———- Predicted text: $[Predicted\ Point]$ Reference text: $[GT\ Point]$