Scenarios and Approaches for Situated Natural Language Explanations

Method

- Prompt ablation study: ‘Base’ — ‘1AD2T3T’
- Meta prompt: use LLM itself to generate prompt
- In-context learning prompt: use other situations’ context let LLM to learn from it.

Prompt Method	Prompt Content
Base	{explanandum} because
1A2F3F	Following is an explanation towards {audience}. {explanandum} because
1A2F3T	Following is an explanation towards {audience}: {explanandum}.
1A2T3F	You are a helpful assistant explaining to {audience}. {explanandum} because
1A2T3T	You are a helpful assistant explaining to {audience}. {explanandum}.
1D2F3F	Following is an explanation about {reason}. {explanandum} because
1D2F3T	Following is an explanation about {reason}. {explanandum}.
1D2T3F	You are a helpful assistant explaining about {desired feature}. {explanandum} because
1D2T3T	You are a helpful assistant explaining about {desired feature}. {explanandum}.
1AD2F3F	Following is an explanation towards {audience}, about {desired feature}. {explanandum} because
1AD2F3T	Following is an explanation towards {audience}, about {desired feature}. {explanandum}
1AD2T3F	You are a helpful assistant explaining to {audience}, about {desired feature}. {explanandum} because
1AD2T3T	You are a helpful assistant explaining to {audience}, about {desired feature}. {explanandum}
Meta prompt	You are a helpful assistant helping me write a prompt. I want to write a prompt to generate an explanation about why {explanandum} to {audience}, about {desired feature}. Give me the prompt directly.
ICL prompt	"For audience_1: Q: Following is an explanation towards {audience2}, about {desired feature2}. {explanandum} because A: {explanation2} Q: Following is an explanation towards {audience3}, about {desired feature3}. {explanandum} because A: {explanation3} Q: Following is an explanation towards {audience1}, about {desired feature1}. {explanandum} because A:"

Scenarios and Approaches for Situated Natural Language Explanations

Evaluation

- Sentence similarity
  - Calculate similarity between ground truth explanation (h_c) and LLM generated (e_j) explanation
- Cross-entropy loss -> matching score
  - Convert similarities to probabilities, then calculate loss

$$p_{cj} = \frac{\exp(\text{sim}(h_c, e_j))}{\sum_{c=1}^{3} \exp(\text{sim}(h_c, e_j))}$$

$$\text{Matching}_j = -\sum_{c=1}^{3} y_c \log(p_{cj})$$