FACEBOOK DATASET PROMPT FOR LLAVA-13b-16bit model

Note add USER: <image>\n to every prompt at start

——————————————————————————————VANILA——

```
# max_length = 100
        prompt = [
        "Classify the meme as hateful or not-hateful. Provide the answer as either hateful or "
        "not-hateful only.\n"
        "Example output for hateful meme : hateful\n"
        "Example output for not-hateful meme : not-hateful\nAssistant:"
         "Assistant: "
       ]
```

——OCR——

```
#max_length=300
prompt = [
      "Classify the above meme as hateful or not-hateful considering the image as well as "
      "the extracted text from the image which is delimited by three backticks.\n"
      f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
      "Provide your answer in the format: hateful or not-hateful.\n"
      "Example output for hateful meme : hateful.\n"
      "Example output for not-hateful meme : not-hateful.\n"
      "Assistant: "
      ]
```

——DEFINITION——

```
#max_length=300
      prompt = [
      "Consider the following definitions.\n"
      "1. hateful - a direct or indirect attack on people based on characteristics, including ethnicity, "
      "race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, "
      "and disability or disease. Attack is defined as violent or dehumanizing (comparing people to non-human "
      "things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking "
      "hate crime is also considered hateful.\n"
      "2. not-hateful - an attack which is not hateful and follows social norms.\n"
      "Classify the above meme as hateful or not-hateful based on the above definitions considering the image "
      "Provide the answer as either hateful or not-hateful only.\n"
      "Example output for hateful meme : hateful\n"
      "Example output for not-hateful meme : not-hateful\n"
       "Assistant: "
      ]
```

——OCR+DEFINITION——

```
#max_length=500
       prompt = [
      "Consider the following definitions.\n"
      "1. hateful - a direct or indirect attack on people based on characteristics, including ethnicity, "
```

```
    "race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, "
    "and disability or disease. Attack is defined as violent or dehumanizing (comparing people to non-human "
    "things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking "
    "hate crime is also considered hateful.\n"
    "2. not-hateful - an attack which is not hateful and follows social norms.\n"
    "Classify the above meme as hateful or not-hateful based on the above definitions considering the image "
    f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
    "Provide the answer as either hateful or not-hateful only.\n"
    "Example output for hateful meme : hateful\n"
    "Example output for not-hateful meme : not-hateful\n"
     "Assistant: "
    ]
```

—----------VANILLA+EXPLANATION—----------

```
    #max_length=400
    prompt = [
    "Classify the meme as hateful or not-hateful.Provide the answer as either hateful or "
    "not-hateful only with an explanation within 30 words explaining your classification.\n"
    "Example output for hateful meme : hateful - Explain within 30 words that why you classified this "
    "meme as hateful.\n"
    "Example output for not-hateful meme : not-hateful - Explain within 30 words that why you classified "
    "this meme as not-hateful.\n"
    "Assistant: "
    ]
```

—----------DEFINITION+EXPLANATION—----------

```
#max_length=600
prompt = [
    "Consider the following definitions.\n"
    "1. hateful - a direct or indirect attack on people based on characteristics, including ethnicity, "
    "race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, "
    "and disability or disease. Attack is defined as violent or dehumanizing (comparing people to non-human "
    "things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking "
    "hate crime is also considered hateful.\n"
    "2. not-hateful - an attack which is not hateful and follows social norms.\n"
    "Classify the above meme as hateful or not-hateful based on the above definitions considering the image "
    "Provide your answer as either hateful or not-hateful only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for hateful meme : hateful - Explain within 30 words that why you classified this "
    "meme as hateful.\n"
    "Example output for not-hateful meme : not-hateful - Explain within 30 words that why you classified "
    "this meme as not-hateful.\n"
    "Assistant: "
    ]
```

—----------OCR+EXPLANATION—----------

```
#max_length=600
prompt = [
    "Classify the above meme as hateful or not-hateful considering the image as well as "
    "the extracted text from the image which is delimited by three backticks.\n"
    f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
    "Provide your answer as either hateful or not-hateful only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for hateful meme : hateful - Explain within 30 words that why you classified this "
    "meme as hateful.\n"
    "Example output for not-hateful meme : not-hateful - Explain within 30 words that why you classified "
    "this meme as not-hateful.\n"
    "Assistant: "
    ]
```

—----------OCR+DEFINITION+EXPLANATION—----------

```
#max_length=800
    prompt = [
    "Consider the following definitions.\n"
    "1. hateful - a direct or indirect attack on people based on characteristics, including ethnicity, "
    "race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, "
    "and disability or disease. Attack is defined as violent or dehumanizing (comparing people to non-human "
    "things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking "
    "hate crime is also considered hateful.\n"
    "2. not-hateful - an attack which is not hateful and follows social norms.\n"
    "Classify the above meme as hateful or not-hateful based on the above definitions considering the image "
    f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
    "Provide your answer as either hateful or not-hateful only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for hateful meme : hateful - Explain within 30 words that why you classified this "
    "meme as hateful.\n"
    "Example output for not-hateful meme : not-hateful - Explain within 30 words that why you classified "
    "this meme as not-hateful.\n"
    "Assistant: "
    ]
```

—----------

HARM C DATASET PROMPT FOR LLAVA-13b-16bit model

Note add USER: <image>\n to every prompt at start

——————————————————————————————VANILA——

```
# max_length = 100
        prompt = [
        "Classify the meme as harmful or not-harmful. Provide the answer as either harmful or "
        "not-harmful only.\n"
        "Example output for harmful meme : harmful\n"
        "Example output for not-harmful meme : not-harmful \nAssistant:"

       ]
```

——OCR——

```
#max_length=300
prompt = [
      "Classify the above meme as harmful or not-harmful considering the image as well as "
      "the extracted text from the image which is delimited by three backticks.\n"
      f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
      "Provide your answer in the format: harmful or not-harmful.\n"
      "Example output for harmful meme : harmful.\n"
      "Example output for not-harmful meme : not-harmful.\n"
      "Assistant: "
      ]
```

——DEFINITION——

```
#max_length=300
      prompt = [
      "Consider the following definitions.\n"
      "1. definition of harmful meme: ```multimodal units consisting of an image and a piece of text embedded"
      "that has the potential to cause harm to an individual, an organization, a community, or the society more"
      "generally. Here, harm includes mental abuse, defamation, psycho-physiological injury, proprietary"
      "damage, emotional disturbance, and compensated public image.```\n"
      "2. definition of not-harmful meme: ```multimodal units consisting of an image and a piece of text embedded
      "which does not cause any harm to an individual, an organization, a community, or the society more generally.``` "
      "Provide the answer as either harmful or not-harmful only.\n"
      "Example output for harmful meme : harmful\n"
      "Example output for not-harmful meme : not-harmful\n"
       "Assistant: "
      ]
```

——OCR+DEFINITION——

```
#max_length=500
      prompt = [
      "Consider the following definitions.\n"
      "1. definition of harmful meme: ```multimodal units consisting of an image and a piece of text embedded"
      "that has the potential to cause harm to an individual, an organization, a community, or the society more"
```

```
"generally. Here, harm includes mental abuse, defamation, psycho-physiological injury, proprietary"
"damage, emotional disturbance, and compensated public image.```\n"
"2. definition of not-harmful meme: ```multimodal units consisting of an image and a piece of text embedded
"which does not cause any harm to an individual, an organization, a community, or the society more generally.``` "
"Classify the above meme as harmful or not-harmful based on the above definitions considering the image "
f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
"Provide the answer as either harmful or not-harmful only.\n"
"Example output for harmful meme : harmful\n"
"Example output for not-harmful meme : not-harmful\n"
 "Assistant: "
]
```

—----------VANILLA+EXPLANATION—----------

```
#max_length=400
prompt = [
"Classify the meme as harmful or not-harmful.Provide the answer as either harmful or "
"not-harmful only with an explanation within 30 words explaining your classification.\n"
"Example output for harmful meme : harmful - Explain within 30 words that why you classified this "
"meme as harmful.\n"
"Example output for not-harmful meme : not-harmful - Explain within 30 words that why you classified "
"this meme as not-harmful.\n"
"Assistant: "
]
```

—----------DEFINITION+EXPLANATION—----------

```
#max_length=600
prompt = [
"Consider the following definitions.\n"
 "1. definition of harmful meme: ```multimodal units consisting of an image and a piece of text embedded"
"that has the potential to cause harm to an individual, an organization, a community, or the society more"
"generally. Here, harm includes mental abuse, defamation, psycho-physiological injury, proprietary"
"damage, emotional disturbance, and compensated public image.```\n"
"2. definition of not-harmful meme: ```multimodal units consisting of an image and a piece of text embedded
"which does not cause any harm to an individual, an organization, a community, or the society more generally.``` "
"Classify the above meme as harmful or not-harmful based on the above definitions considering the image "
"Provide your answer as either harmful or not-harmful only with an explanation within 30 words explaining "
"your classification.\n"
"Example output for harmful meme : harmful - Explain within 30 words that why you classified this "
"meme as harmful.\n"
"Example output for not-harmful meme : not-harmful - Explain within 30 words that why you classified "
"this meme as not-harmful.\n"
"Assistant: "
]
```

—----------OCR+EXPLANATION—----------

```
#max_length=600
prompt = [
      "Classify the above meme as harmful or not-harmful considering the image as well as "
      "the extracted text from the image which is delimited by three backticks.\n"
      f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
      "Provide your answer as either harmful or not-harmful only with an explanation within 30 words explaining "
      "your classification.\n"
      "Example output for harmful meme : harmful - Explain within 30 words that why you classified this "
      "meme as harmful.\n"
      "Example output for not-harmful meme : not-harmful - Explain within 30 words that why you classified "
      "this meme as not-harmful.\n"
      "Assistant: "
      ]
```

—----------OCR+DEFINITION+EXPLANATION—----------

```
#max_length=800
      prompt = [
      "Consider the following definitions.\n"
      "1. definition of harmful meme: ```multimodal units consisting of an image and a piece of text embedded"
      "that has the potential to cause harm to an individual, an organization, a community, or the society more"
      "generally. Here, harm includes mental abuse, defamation, psycho-physiological injury, proprietary"
      "damage, emotional disturbance, and compensated public image.```\n"
      "2. definition of not-harmful meme: ```multimodal units consisting of an image and a piece of text embedded
      "which does not cause any harm to an individual, an organization, a community, or the society more generally.``` "
      "Classify the above meme as harmful or not-harmful based on the above definitions considering the image "
      f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
      "Provide your answer as either harmful or not-harmful only with an explanation within 30 words explaining "
      "your classification.\n"
      "Example output for harmful meme : harmful - Explain within 30 words that why you classified this "
      "meme as harmful.\n"
      "Example output for not-harmful meme : not-harmful - Explain within 30 words that why you classified "
      "this meme as not-harmful.\n"
       "Assistant: "
      ]
```

—----------

MAMI DATASET PROMPT FOR LLAVA-13b-16bit model

Note add USER: <image>\n to every prompt at start

——————————————————————————————VANILA——

```
# max_length = 100
        prompt = [
        "Classify the meme as misogynistic or not-misogynistic. Provide the answer as either misogynistic or "
        "not-misogynistic only.\n"
        "Example output for misogynistic meme : misogynistic\n"
        "Example output for not-misogynistic meme : not-misogynistic \nAssistant:"

       ]
```

——OCR——

```
#max_length=300
prompt = [
      "Classify the above meme as misogynistic or not-misogynistic considering the image as well as "
      "the extracted text from the image which is delimited by three backticks.\n"
      f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
      "Provide your answer in the format: misogynistic or not-misogynistic.\n"
      "Example output for misogynistic meme : misogynistic.\n"
      "Example output for not-misogynistic meme : not-misogynistic.\n"
      "Assistant: "
      ]
```

——DEFINITION——

```
#max_length=300
      prompt = [
      "Consider the following definitions.\n"
      "1. definition of 'misogynistic' meme: ```a meme is misogynous if it conceptually describes an offensive, "
      "sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group "
      "of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.```"
      "2. definition of 'not-misogynistic' meme: ```a meme that does not express any form "
      "of hate against women.```"
      "Provide the answer as either misogynistic or not-misogynistic only.\n"
      "Example output for misogynistic meme : misogynistic\n"
      "Example output for not-misogynistic meme : not-misogynistic\n"
       "Assistant: "
      ]
```

——OCR+DEFINITION——

```
#max_length=500
      prompt = [
      "Consider the following definitions.\n"
      "1. definition of 'misogynistic' meme: ```a meme is misogynous if it conceptually describes an offensive, "
```

```
    "sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group "
    "of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.```"
    "2. definition of 'not-misogynistic' meme: ```a meme that does not express any form "
    "of hate against women.```"
    "Classify the above meme as misogynistic or not-misogynistic based on the above definitions considering the image "
    f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
    "Provide the answer as either misogynistic or not-misogynistic only.\n"
    "Example output for misogynistic meme : misogynistic\n"
    "Example output for not-misogynistic meme : not-misogynistic\n"
     "Assistant: "
    ]
```

—----------VANILLA+EXPLANATION—----------

```
    #max_length=400
     prompt = [
    "Classify the meme as misogynistic or not-misogynistic.Provide the answer as either misogynistic or "
    "not-misogynistic only with an explanation within 30 words explaining your classification.\n"
    "Example output for misogynistic meme : misogynistic - Explain within 30 words that why you classified this "
    "meme as misogynistic.\n"
    "Example output for not-misogynistic meme : not-misogynistic - Explain within 30 words that why you classified "
    "this meme as not-misogynistic.\n"
    "Assistant: "
    ]
```

—----------DEFINITION+EXPLANATION—----------

```
#max_length=600
prompt = [
    "Consider the following definitions.\n"
    "1. definition of 'misogynistic' meme: ```a meme is misogynous if it conceptually describes an offensive, "
    "sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group "
    "of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.```"
    "2. definition of 'not-misogynistic' meme: ```a meme that does not express any form "
    "of hate against women.```"
    "Classify the above meme as misogynistic or not-misogynistic based on the above definitions considering the image "
    "Provide your answer as either misogynistic or not-misogynistic only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for misogynistic meme : misogynistic - Explain within 30 words that why you classified this "
    "meme as misogynistic.\n"
    "Example output for not-misogynistic meme : not-misogynistic - Explain within 30 words that why you classified "
    "this meme as not-misogynistic.\n"
    "Assistant: "
    ]
```

----------OCR+EXPLANATION----------

```
#max_length=600
prompt = [
    "Classify the above meme as misogynistic or not-misogynistic considering the image as well as "
    "the extracted text from the image which is delimited by three backticks.\n"
    f"""Extracted text from the image: ```{image_metadata['text']}```\n"""
    "Provide your answer as either misogynistic or not-misogynistic only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for misogynistic meme : misogynistic - Explain within 30 words that why you classified this "
    "meme as misogynistic.\n"
    "Example output for not-misogynistic meme : not-misogynistic - Explain within 30 words that why you classified "
    "this meme as not-misogynistic.\n"
    "Assistant: "
    ]
```

----------OCR+DEFINITION+EXPLANATION----------

```
#max_length=800
    prompt = [
    "Consider the following definitions.\n"
    "1. definition of 'misogynistic' meme: ```a meme is misogynous if it conceptually describes an offensive, "
    "sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group "
    "of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.```"
    "2. definition of 'not-misogynistic' meme: ```a meme that does not express any form "
    "of hate against women.```"
    "Classify the above meme as misogynistic or not-misogynistic based on the above definitions considering the image "
    f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
    "Provide your answer as either misogynistic or not-misogynistic only with an explanation within 30 words explaining "
    "your classification.\n"
    "Example output for misogynistic meme : misogynistic - Explain within 30 words that why you classified this "
    "meme as misogynistic.\n"
    "Example output for not-misogynistic meme : not-misogynistic - Explain within 30 words that why you classified "
    "this meme as not-misogynistic.\n"
     "Assistant: "
    ]
```

TRY 1

```
#MAX LENGTH = 800
"USER: <image>\nConsider the following definitions.\n"
"1. definition of 'misogynistic' meme: ```a meme is misogynous if it conceptually describes an offensive, "
"sexist or hateful scene (weak or strong, implicitly or explicitly) having as target a woman or a group "
"of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.```"
"2. definition of 'not-misogynistic' meme: ```a meme that does not express any form "
"of hate against women.```"
"Classify the above meme as misogynistic or not-misogynistic based on the above definitions considering the image "
f"as well as the extracted text from the image delimited by three backticks```{image_metadata['text']}```.\n"
"Provide the answer as either misogynistic or not-misogynistic only.\n"
"Example output for misogynistic meme : misogynistic\n"
"Example output for not-misogynistic meme : not-misogynistic\n"
    "\nASSISTANT: First, I will thoroughly understand the provided definitions of misogynistic content. Then, I will examine the image and any text obtained through OCR to determine if any misogynistic elements are present. If 'Yes', I will classify the meme as misogynistic; otherwise, I will classify it as not misogynistic. My output will be preceded by the message 'The meme is':"
```