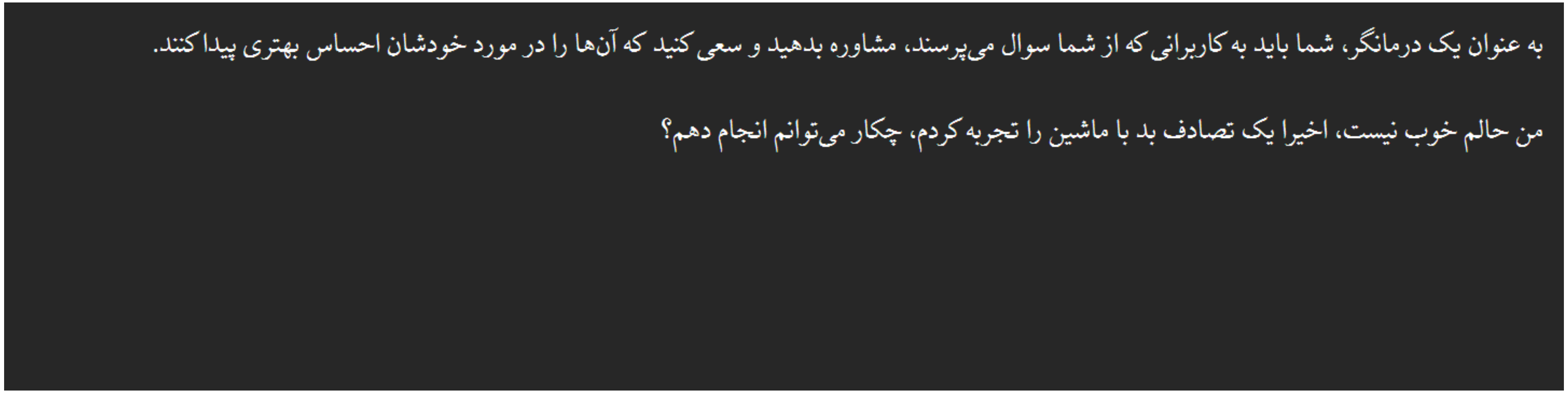
System-User Prompt Design

System-User Prompt Design

- One way to increase the performance of the model is to use System-User Prompt structure.
- In this way, the system prompt is constructed around defining the task and what model should generate.
- The user prompt is considered as the prompt given by the user.
- The system prompt is mostly unchanged throughout the iterations and user prompt is always dynamic.

System-User Prompt Design

- The input prompt which is only in Persian language:



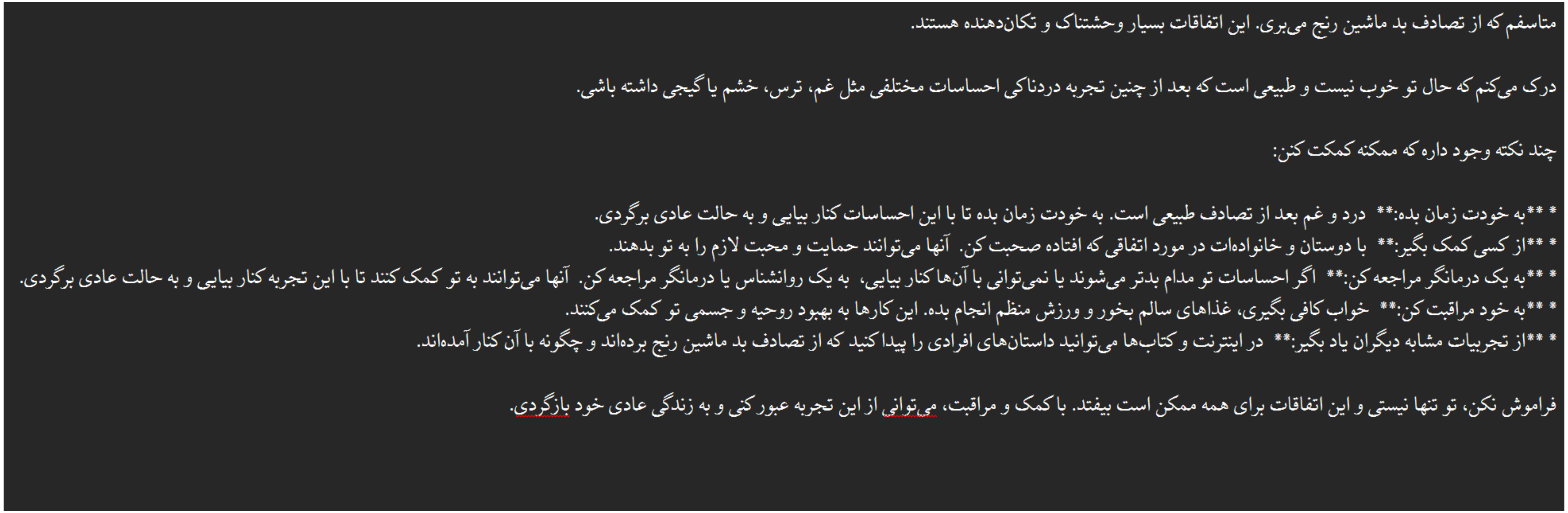
به عنوان یک درمانگر، شما باید به کاربرانی که از شما سوال می‌پرسند، مشاوره بدهید و سعی کنید که آن‌ها را در مورد خودشان احساس بهتری پیدا کنند.

من حالم خوب نیست، اخیرا یک تصادف بد با ماشین را تجربه کردم، چکار می‌توانم انجام دهم؟

- Without the use of any system-user prompt instruction.

System-User Prompt Design

- And this is the result form Gemma2 9b it:

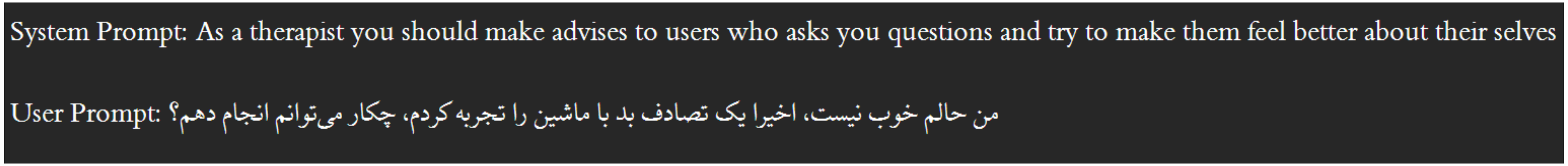


```
متاسفم که از تصادف بد ماشین رنج می‌بری. این اتفاقات بسیار وحشتناک و تکان‌دهنده هستند.

درک می‌کنم که حال تو خوب نیست و طبیعی است که بعد از چنین تجربه دردناکی احساسات مختلفی مثل غم، ترس، خشم یا گیجی داشته باشی.

چند نکته وجود داره که ممکنه کمکت کنن:

* **به خودت زمان بده:**  درد و غم بعد از تصادف طبیعی است. به خودت زمان بده تا با این احساسات کنار بیایی و به حالت عادی برگردی.
* **از کسی کمک بگیر:**  با دوستان و خانواده‌ات در مورد اتفاقی که افتاده صحبت کن.  آنها می‌توانند حمایت و محبت لازم را به تو بدهند.
* **به یک درمانگر مراجعه کن:**  اگر احساسات تو مدام بدتر می‌شوند یا نمی‌توانی با آن‌ها کنار بیایی،  به یک روانشناس یا درمانگر مراجعه کن.  آنها می‌توانند به تو کمک کنند تا با این تجربه کنار بیایی و به حالت عادی برگردی.
* **به خود مراقبت کن:**  خواب کافی بگیری، غذاهای سالم بخور و ورزش منظم انجام بده. این کارها به بهبود روحیه و جسمی تو کمک می‌کنند.
* **از تجربیات مشابه دیگران یاد بگیر:**  در اینترنت و کتاب‌ها می‌توانید داستان‌های افرادی را پیدا کنید که از تصادف بد ماشین رنج برده‌اند و چگونه با آن کنار آمده‌اند.

فراموش نکن، تو تنها نیستی و این اتفاقات برای همه ممکن است بیفتد. با کمک و مراقبت، می‌توانی از این تجربه عبور کنی و به زندگی عادی خود بازگردی.
```

- Which is quite acceptable

System-User Prompt Design

- Now we changed the input prompt to have system-user instruction.



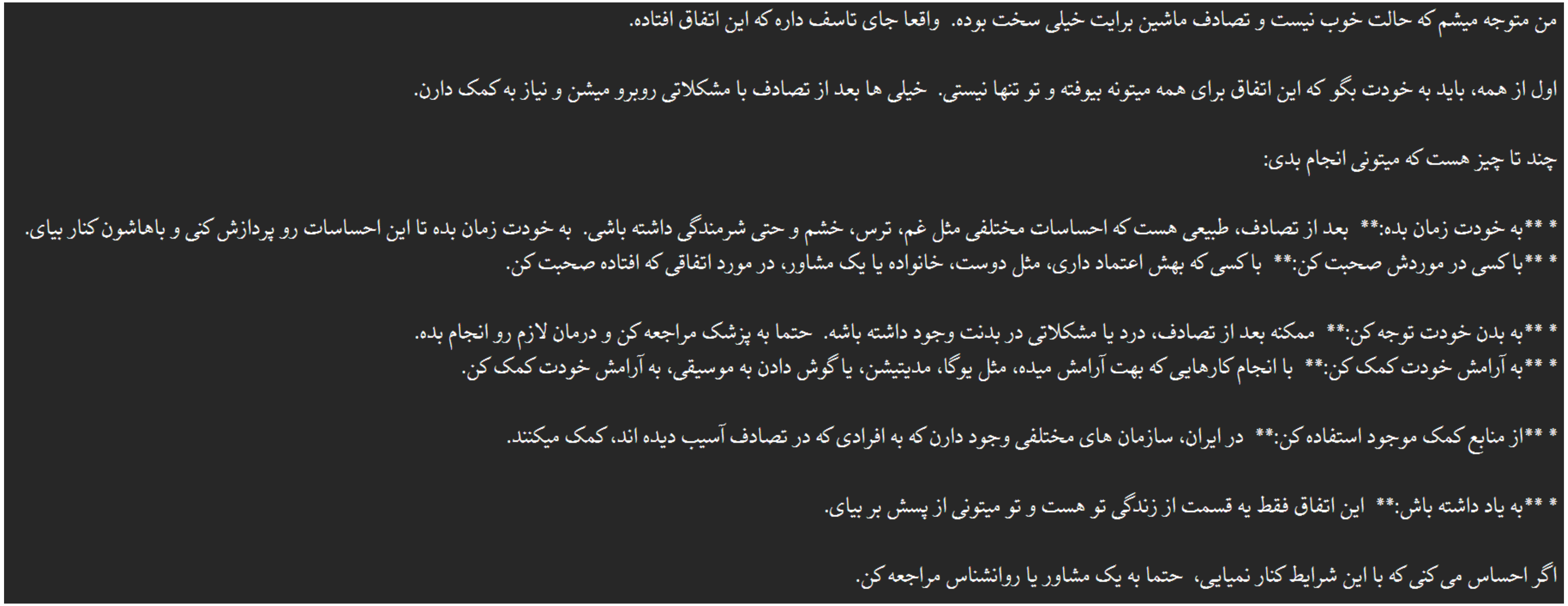
System Prompt: As a therapist you should make advises to users who asks you questions and try to make them feel better about their selves

User Prompt: من حالم خوب نیست، اخیرا یک تصادف بد با ماشین را تجربه کردم، چکار می‌توانم انجام دهم؟

- In a way that System prompt being in English language and the user prompt being in Persian language.
  - This is due to the reason that the model can understand instructions better in English language and have a better reasoning in that situation.

System-User Prompt Design

- This the result form Gemma2 9b it for changed prompt:



```
من متوجه میشم که حالت خوب نیست و تصادف ماشین برایت خیلی سخت بوده.  واقعا جای تاسف داره که این اتفاق افتاده.

اول از همه، باید به خودت بگو که این اتفاق برای همه میتونه بیوفته و تو تنها نیستی.  خیلی ها بعد از تصادف با مشکلاتی روبرو میشن و نیاز به کمک دارن.

چند تا چیز هست که میتونی انجام بدی:

* **به خودت زمان بده:**  بعد از تصادف، طبیعی هست که احساسات مختلفی مثل غم، ترس، خشم و حتی شرمندگی داشته باشی.  به خودت زمان بده تا این احساسات رو پردازش کنی و باهاشون کنار بیای.
* **با کسی در موردش صحبت کن:**  با کسی که بهش اعتماد داری، مثل دوست، خانواده یا یک مشاور، در مورد اتفاقی که افتاده صحبت کن.

* **به بدن خودت توجه کن:**  ممکنه بعد از تصادف، درد یا مشکلاتی در بدنت وجود داشته باشه.  حتما به پزشک مراجعه کن و درمان لازم رو انجام بده.
* **به آرامش خودت کمک کن:**  با انجام کارهایی که بهت آرامش میده، مثل یوگا، مدیتیشن، یا گوش دادن به موسیقی، به آرامش خودت کمک کن.

* **از منابع کمک موجود استفاده کن:**  در ایران، سازمان های مختلفی وجود دارن که به افرادی که در تصادف آسیب دیده اند، کمک میکنند.

* **به یاد داشته باش:**  این اتفاق فقط یه قسمت از زندگی تو هست و تو میتونی از پسش بر بیای.

اگر احساس می کنی که با این شرایط کنار نمیایی،  حتما به یک مشاور یا روانشناس مراجعه کن.
```

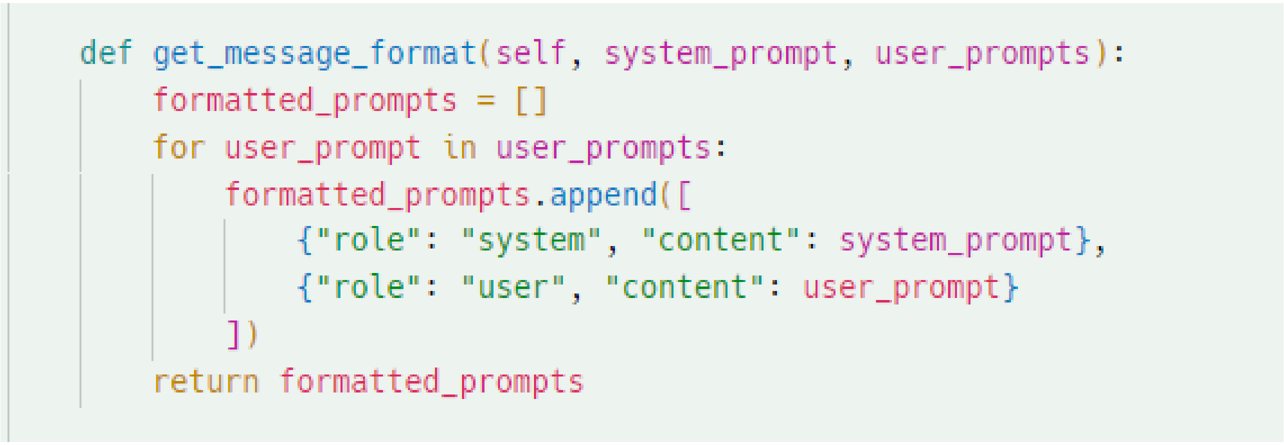
- Not only the reference time was 2x faster, but also the results became more natural and enhanced.

System-User Prompt Design

- As the result of the little experience conducted, it is worthwhile to test the system-user prompt design with system prompt being in English (including the k-shot examples in Farsi) and the user prompt being the same.
- In that case it is feasible to not repeat the instructions due to the better understanding of the model.

System-User Prompt Results

- Now for this part, we changed the code in a way that accepts both roles of ‘user’ and ‘system’.

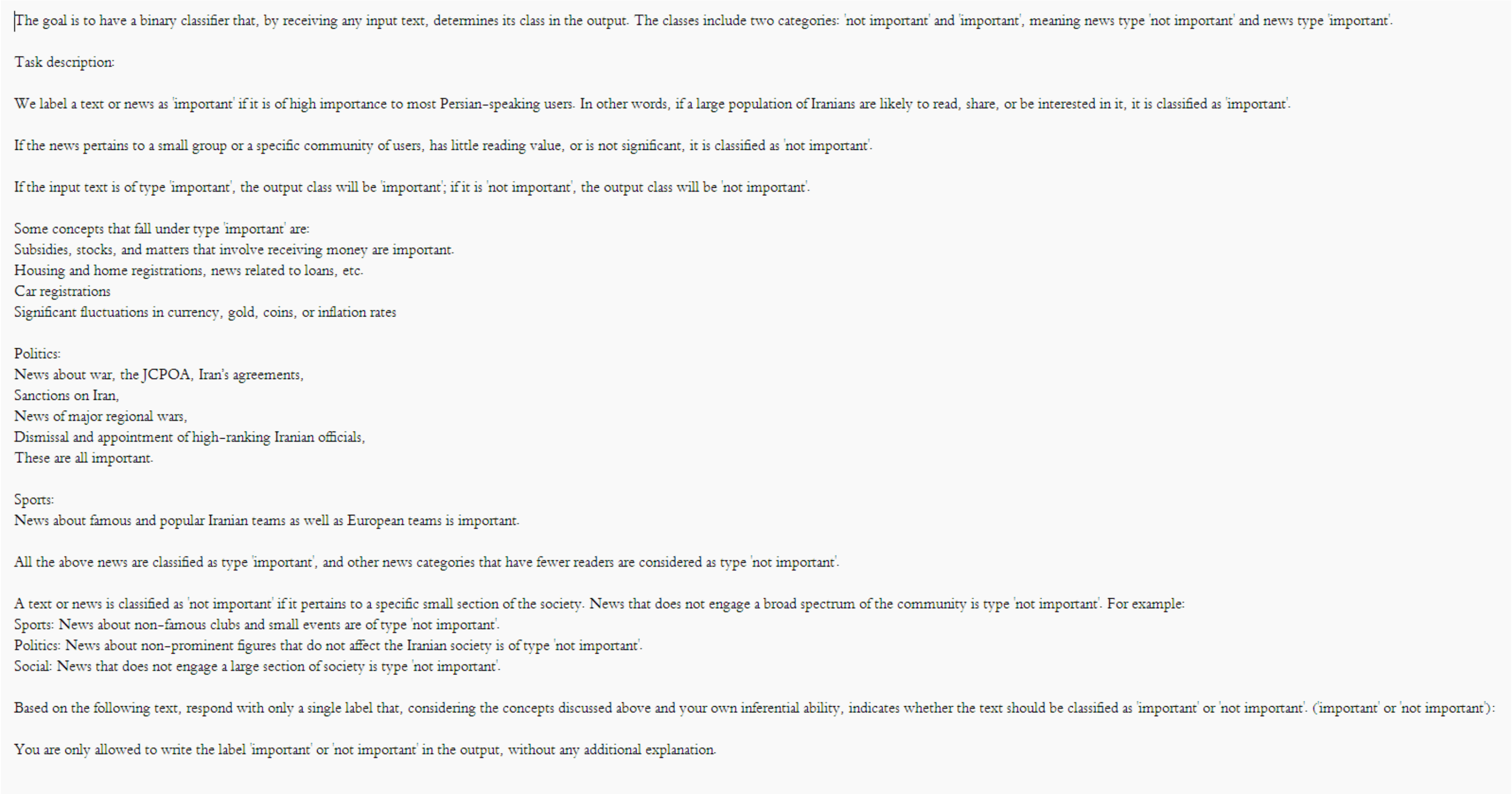


```
def get_message_format(self, system_prompt, user_prompts):
    formatted_prompts = []
    for user_prompt in user_prompts:
        formatted_prompts.append([
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": user_prompt}
        ])
    return formatted_prompts
```

- It should be mentioned that gemma2 doesn’t support any other role than ‘user’.

System-User Prompt Results

- Moreover, we changed the system prompt to English and maintained the input and samples in Persian.



The goal is to have a binary classifier that, by receiving any input text, determines its class in the output. The classes include two categories: 'not important' and 'important', meaning news type 'not important' and news type 'important'.

Task description:

We label a text or news as 'important' if it is of high importance to most Persian-speaking users. In other words, if a large population of Iranians are likely to read, share, or be interested in it, it is classified as 'important'.

If the news pertains to a small group or a specific community of users, has little reading value, or is not significant, it is classified as 'not important'.

If the input text is of type 'important', the output class will be 'important'; if it is 'not important', the output class will be 'not important'.

Some concepts that fall under type 'important' are:
Subsidies, stocks, and matters that involve receiving money are important.
Housing and home registrations, news related to loans, etc.
Car registrations
Significant fluctuations in currency, gold, coins, or inflation rates

Politics:
News about war, the JCPOA, Iran's agreements,
Sanctions on Iran,
News of major regional wars,
Dismissal and appointment of high-ranking Iranian officials,
These are all important.

Sports:
News about famous and popular Iranian teams as well as European teams is important.

All the above news are classified as type 'important', and other news categories that have fewer readers are considered as type 'not important'.

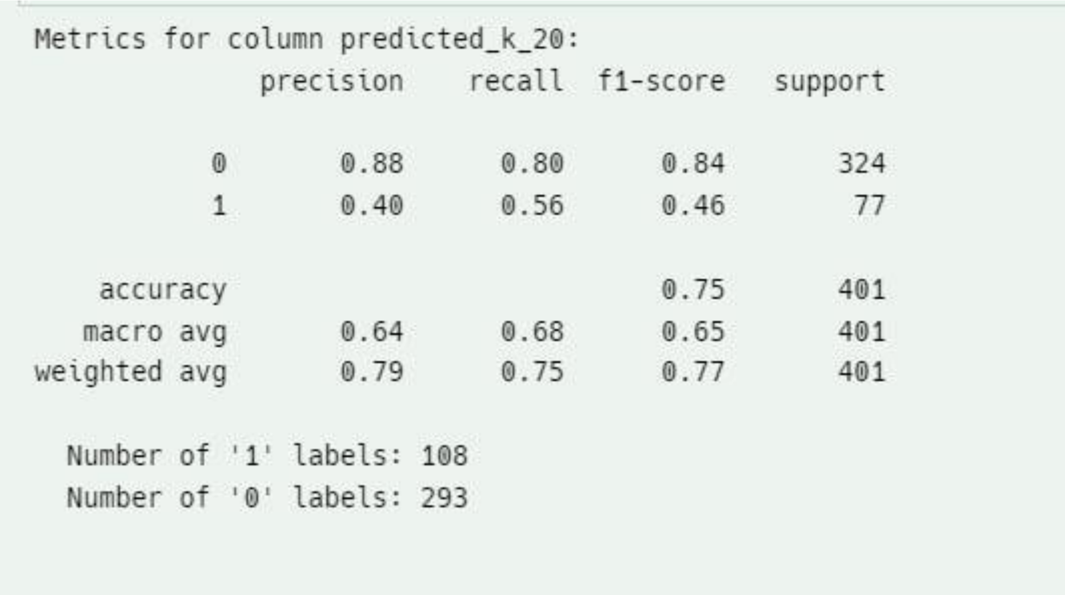
A text or news is classified as 'not important' if it pertains to a specific small section of the society. News that does not engage a broad spectrum of the community is type 'not important'. For example:
Sports: News about non-famous clubs and small events are of type 'not important'.
Politics: News about non-prominent figures that do not affect the Iranian society is of type 'not important'.
Social: News that does not engage a large section of society is type 'not important'.

Based on the following text, respond with only a single label that, considering the concepts discussed above and your own inferential ability, indicates whether the text should be classified as 'important' or 'not important'. ('important' or 'not important'):

You are only allowed to write the label 'important' or 'not important' in the output, without any additional explanation.

System-User Prompt Results

- The shown figure illustrates the result for the first 400 samples with k=20 shot learning and in only title mode.



```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.88      0.80      0.84       324
           1       0.40      0.56      0.46        77

    accuracy                           0.75       401
   macro avg       0.64      0.68      0.65       401
weighted avg       0.79      0.75      0.77       401

  Number of '1' labels: 108
  Number of '0' labels: 293
```

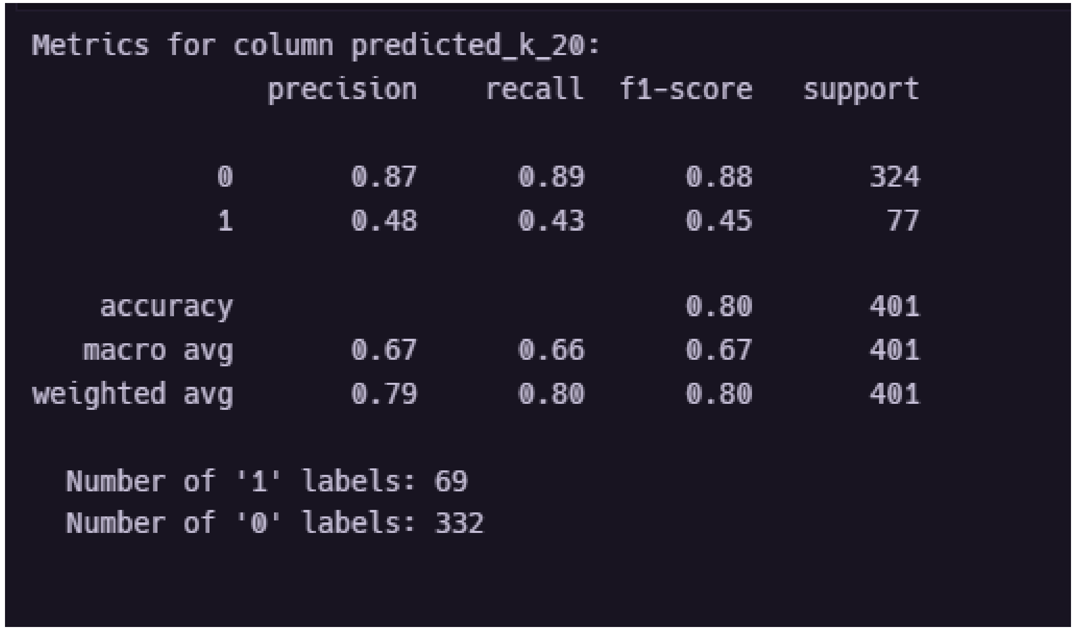
- Here the ‘important’ and ‘not important’ labels are used.

System-User Prompt Results

- In conclusion, it seems that both ‘system-user’ prompt design and using English system prompt result in better performance.
    - It is highly possible that the model has better reasoning power in English rather than Persian.

System-User Prompt Results

- In further step, we changed the labels to ‘0’ and ‘1’s and here is the results.



```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       324
           1       0.48      0.43      0.45        77

    accuracy                           0.80       401
   macro avg       0.67      0.66      0.67       401
weighted avg       0.79      0.80      0.80       401

  Number of '1' labels: 69
  Number of '0' labels: 332
```

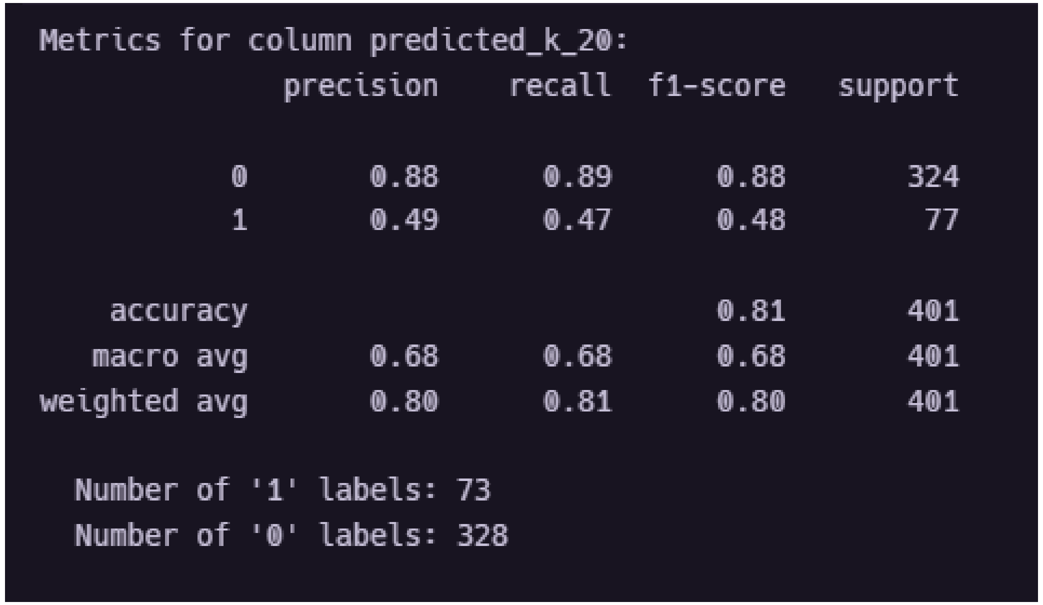
- As we can see there are fewer ‘1’ label predicted hence the overall f1-score saw an increase.

System-User Prompt Results

- Changing labels from ‘important’ and ‘not important’ to ‘1’s and ‘0’s result it 2% of increase in accuracy.
  - The main reason is that in second situation, the model predicted less ‘1’ labels.
- Still, it is arguable what the real reason is behind predicting less ‘1’ labels and why in this case LLM would interpret the prompt in a different way.

System-User Prompt Results

- In next step, we remove the description related to ‘not important’ news. This result in 1% of increase in f1-score.

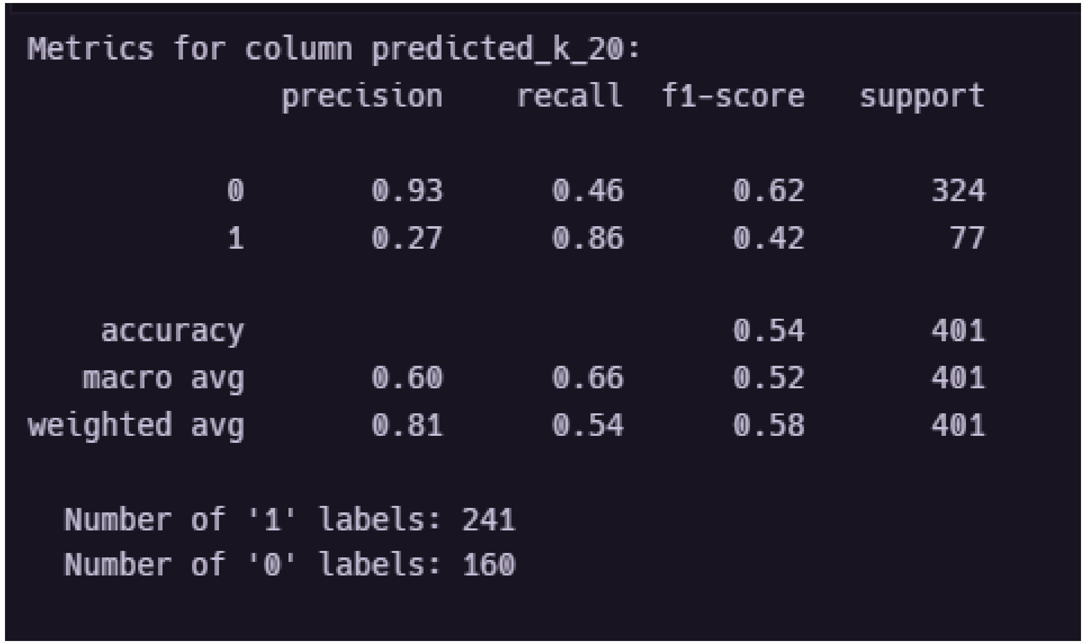


```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.88      0.89      0.88       324
           1       0.49      0.47      0.48        77

    accuracy                           0.81       401
   macro avg       0.68      0.68      0.68       401
weighted avg       0.80      0.81      0.80       401

  Number of '1' labels: 73
  Number of '0' labels: 328
```

- The better result could be because of less additional information in the prompt.

System-User Prompt Results

- Now we tested Persian system prompt, as we can see the f1-score is 16% less than the English prompt.

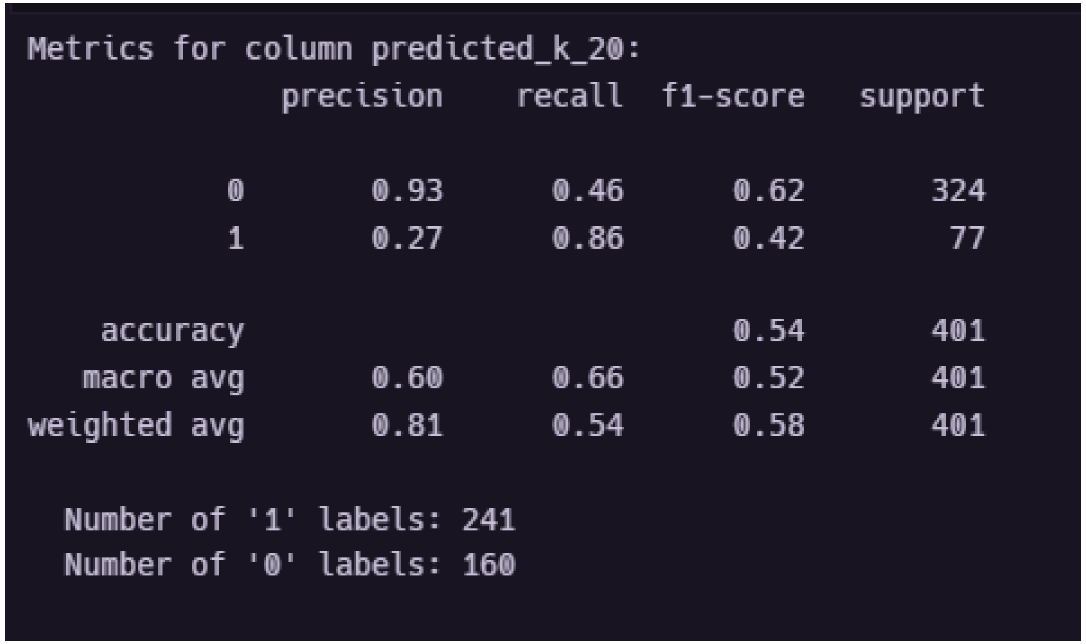


```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.93      0.46      0.62       324
           1       0.27      0.86      0.42        77

    accuracy                           0.54       401
   macro avg       0.60      0.66      0.52       401
weighted avg       0.81      0.54      0.58       401

  Number of '1' labels: 241
  Number of '0' labels: 160
```

- This clearly depicts the fact that LLMs have better reasoning performance in English rather than Farsi.

System-User Prompt Results

- Now we tested Persian system prompt, as we can see the f1-score is 16% less than the English prompt.

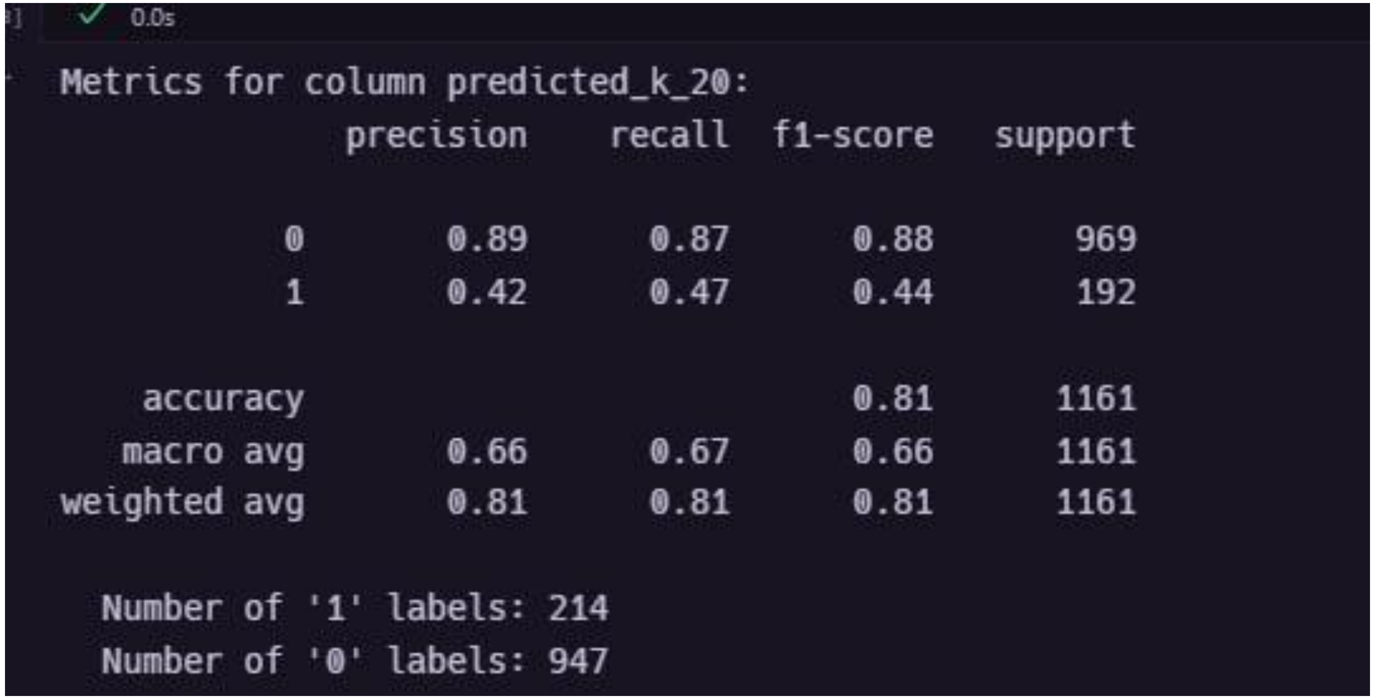


```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.93      0.46      0.62       324
           1       0.27      0.86      0.42        77

    accuracy                           0.54       401
   macro avg       0.60      0.66      0.52       401
weighted avg       0.81      0.54      0.58       401

  Number of '1' labels: 241
  Number of '0' labels: 160
```

- This clearly depicts the fact that LLMs have better reasoning performance in English rather than Farsi.

System-User Prompt Results

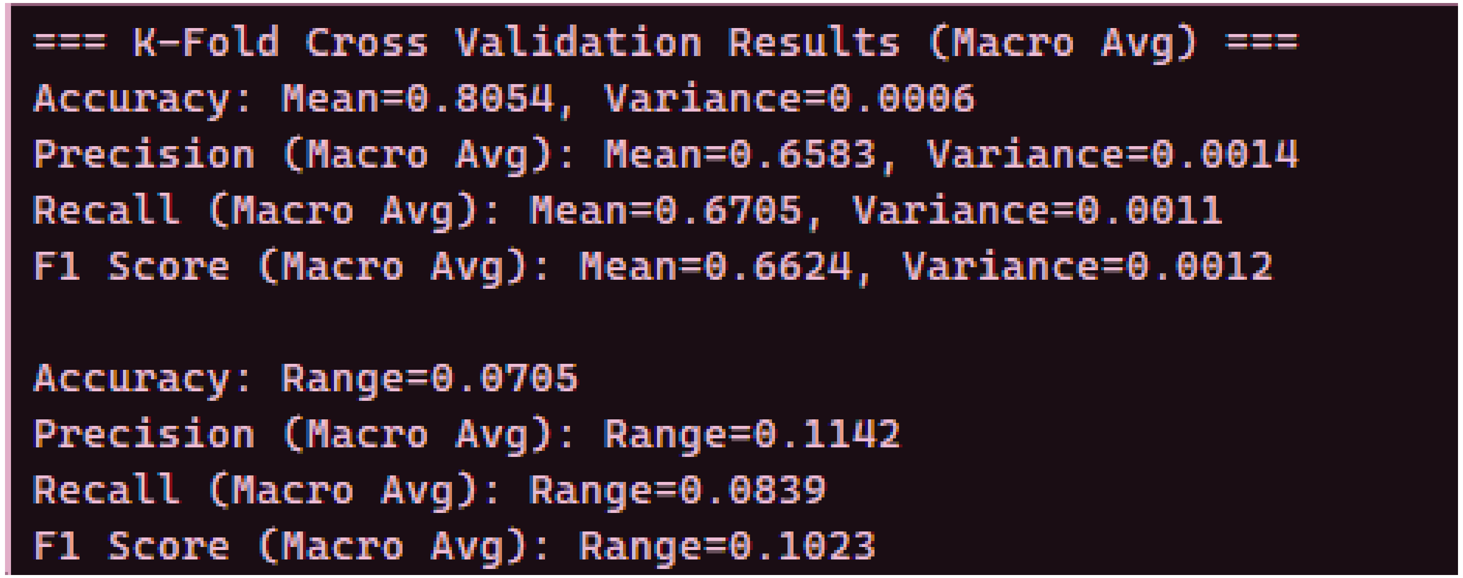
- The result of English system prompt in all test data.



```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.89      0.87      0.88       969
           1       0.42      0.47      0.44       192

    accuracy                           0.81      1161
   macro avg       0.66      0.67      0.66      1161
weighted avg       0.81      0.81      0.81      1161

  Number of '1' labels: 214
  Number of '0' labels: 947
```

System-User Prompt Results

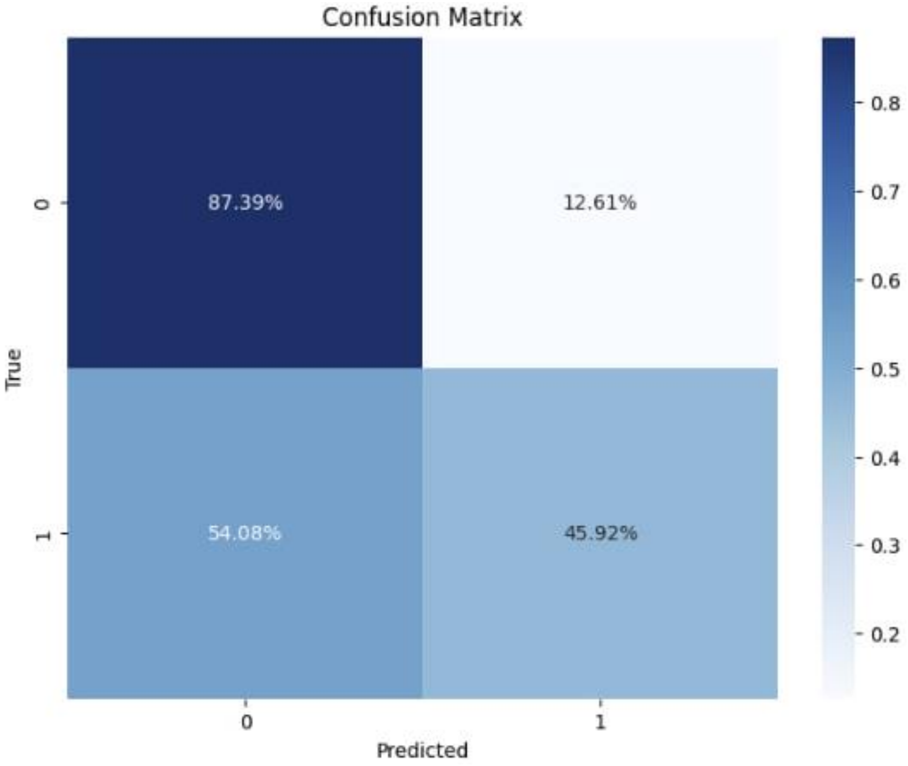
- The Variance in different kfold:



```
=== K-Fold Cross Validation Results (Macro Avg) ===
Accuracy: Mean=0.8054, Variance=0.0006
Precision (Macro Avg): Mean=0.6583, Variance=0.0014
Recall (Macro Avg): Mean=0.6705, Variance=0.0011
F1 Score (Macro Avg): Mean=0.6624, Variance=0.0012

Accuracy: Range=0.0705
Precision (Macro Avg): Range=0.1142
Recall (Macro Avg): Range=0.0839
F1 Score (Macro Avg): Range=0.1023
```

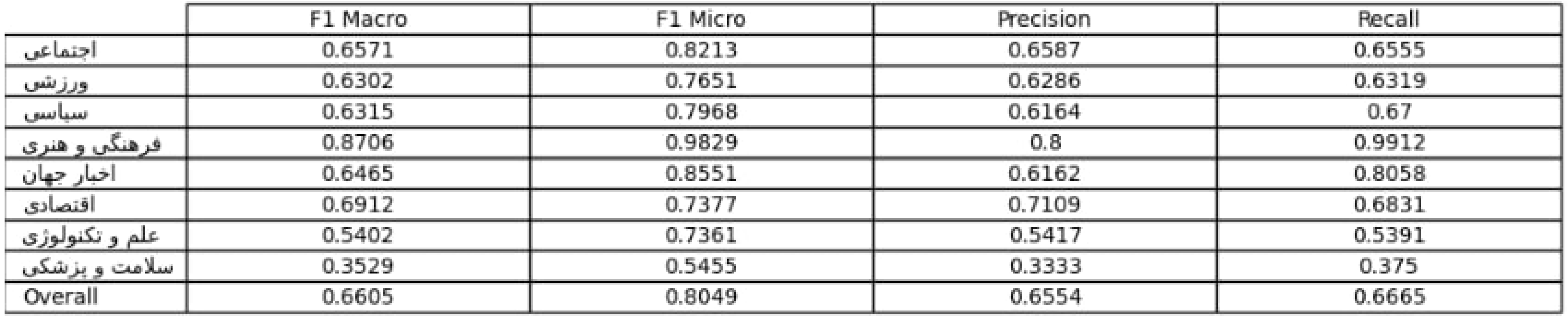
System-User Prompt Results

- The confusion matrix for best result:



System-User Prompt Results

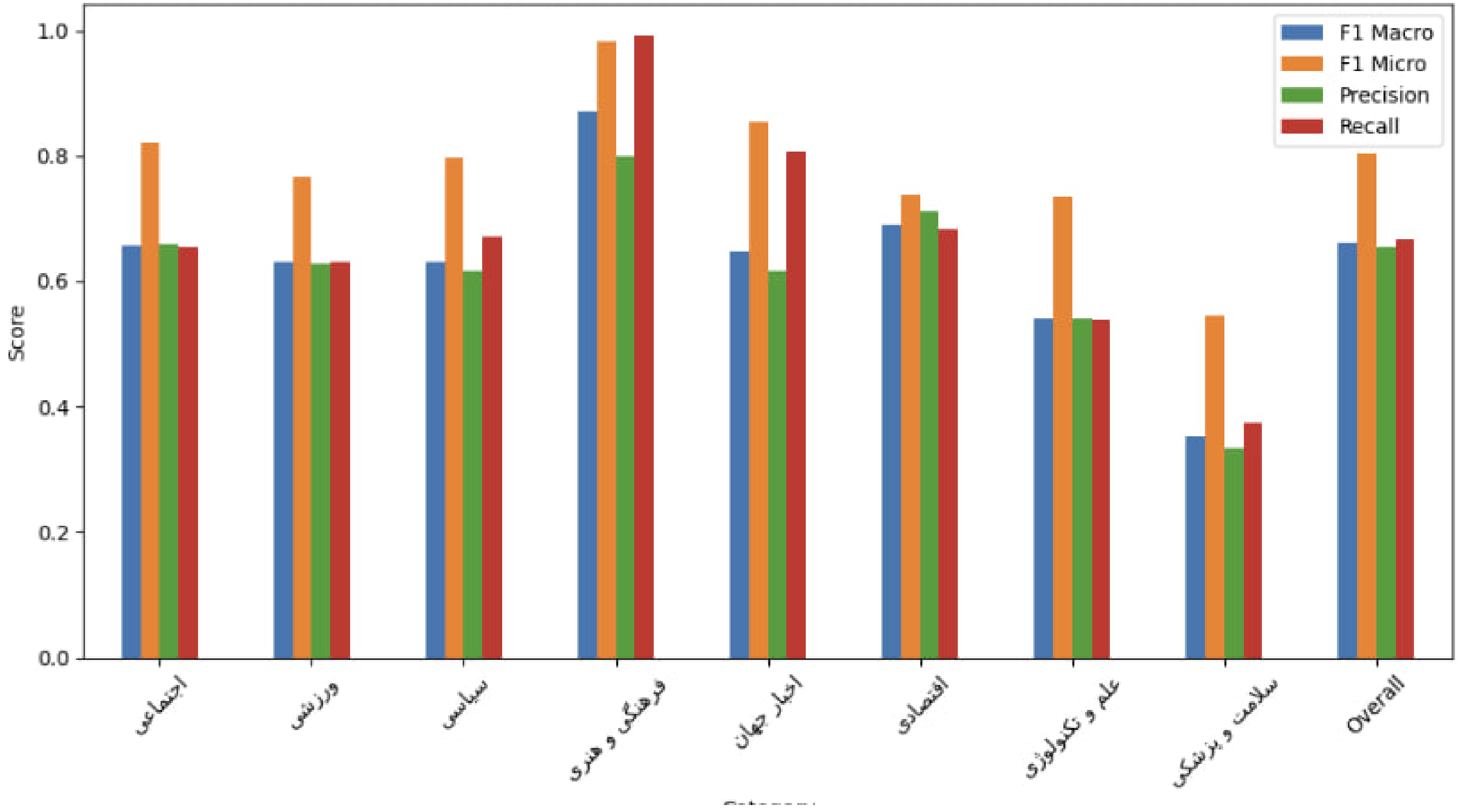
- The f1-score across different categorize:



	F1 Macro	F1 Micro	Precision	Recall
اجتماعی	0.6571	0.8213	0.6587	0.6555
ورزشی	0.6302	0.7651	0.6286	0.6319
سیاسی	0.6315	0.7968	0.6164	0.67
فرهنگی و هنری	0.8706	0.9829	0.8	0.9912
اخبار جهان	0.6465	0.8551	0.6162	0.8058
اقتصادی	0.6912	0.7377	0.7109	0.6831
علم و تکنولوژی	0.5402	0.7361	0.5417	0.5391
سلامت و پزشکی	0.3529	0.5455	0.3333	0.375
Overall	0.6605	0.8049	0.6554	0.6665

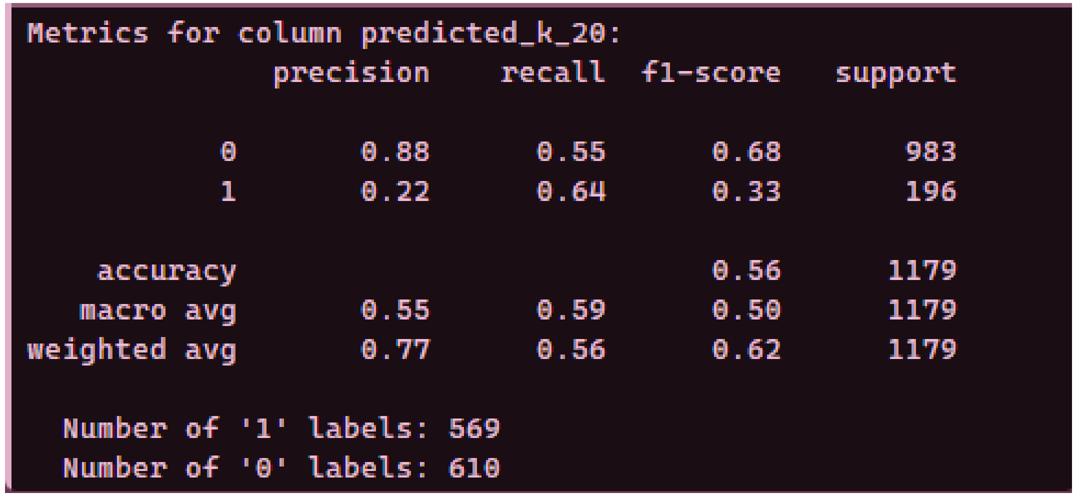
System-User Prompt Results

- The bar chart across different categorize:



System-User Prompt Results

- The result of English system prompt in all test data for full text:

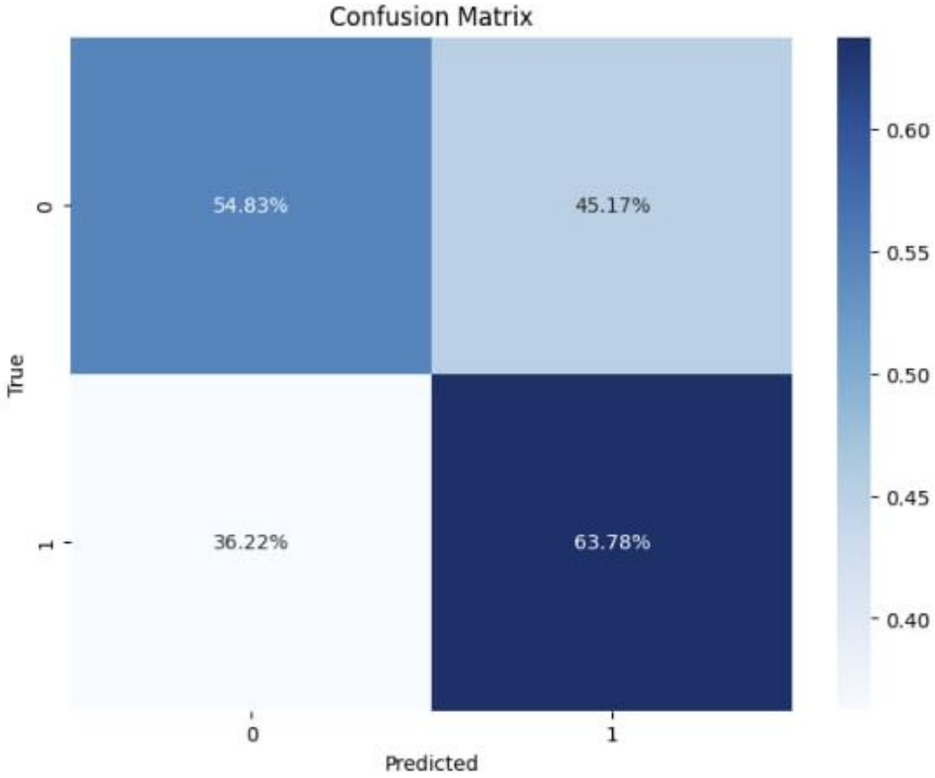


```
Metrics for column predicted_k_20:
              precision    recall  f1-score   support

           0       0.88      0.55      0.68       983
           1       0.22      0.64      0.33       196

    accuracy                           0.56      1179
   macro avg       0.55      0.59      0.50      1179
weighted avg       0.77      0.56      0.62      1179

  Number of '1' labels: 569
  Number of '0' labels: 610
```

- It seems that the user prompt in full text hindered the accuracy in a noticeable way.

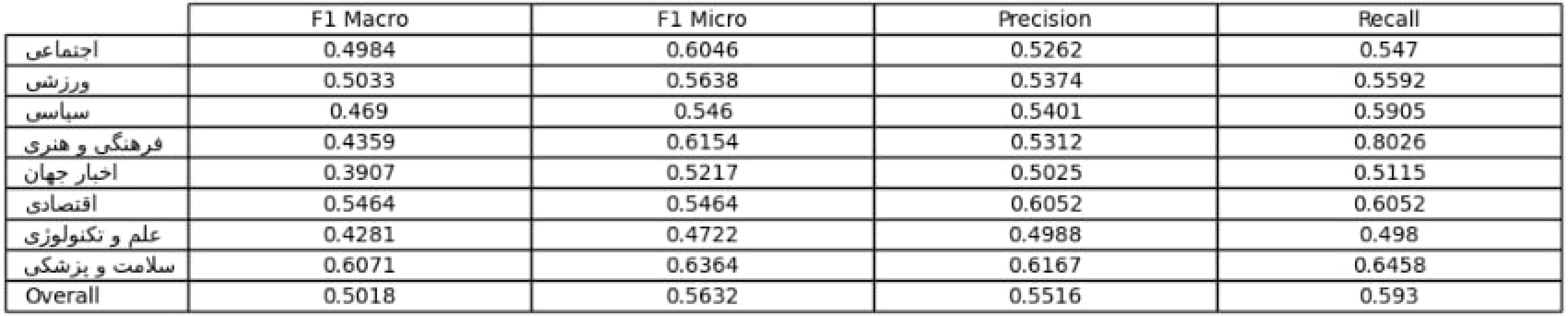
System-User Prompt Results

- The confusion matrix for full text:



System-User Prompt Results

- F1-score across different categories for the full text condition:



	F1 Macro	F1 Micro	Precision	Recall
اجتماعی	0.4984	0.6046	0.5262	0.547
ورزشی	0.5033	0.5638	0.5374	0.5592
سیاسی	0.469	0.546	0.5401	0.5905
فرهنگی و هنری	0.4359	0.6154	0.5312	0.8026
اخبار جهان	0.3907	0.5217	0.5025	0.5115
اقتصادی	0.5464	0.5464	0.6052	0.6052
علم و تکنولوژی	0.4281	0.4722	0.4988	0.498
سلامت و پزشکی	0.6071	0.6364	0.6167	0.6458
Overall	0.5018	0.5632	0.5516	0.593