Philip Georgis
January 31, 2025

IMDB Film Review Sentiment Analysis with SLMs: Extended Study

1. Overview

In the present study extension, several additional prompts and prompting techniques are presented and evaluated using the same two Qwen2.5 SLMs on the IMDB film review dataset for sentiment analysis as in the original study. The structure of the chain-of-thought (CoT) and few-shot prompts presented in the original study were revised and retested, and a more advanced few-shot prompt that leverages dynamic example selection was implemented and tested. Three further prompting techniques using a quantitative approach to sentiment analysis combined with self-consistency and/or CoT prompting were likewise implemented and evaluated on the IMDB movie review dataset.

2. New Methods

a. Dynamic Few-Shot with Most Similar Examples

The few-shot prompt from the original study was enhanced through the implementation of dynamic example selection for inclusion in the prompt. To enable this approach, a subset of the train partition of the IMDB dataset was extracted and embedded into a 384-dimensional vector space using the all-MiniLM-L6-v2 sentence embedding model from HuggingFace in advance and stored in a FAISS vector index. During inference, rather than statically including the same N examples per sentiment classification query, the review to be classified is embedded into the same vector space and the N most similar positive and negative reviews from the pool of embedded training set reviews are retrieved as examples. This approach was motivated by the hypothesis that showing more semantically similar examples to the film review to be classified together with their ground-truth labels might aid the SLM and improve performance versus showing randomly selected reviews as examples. The subset of reviews from the training set included in the embedded example pool was limited to those containing between 75 and 150 words, as embedding similarities are impacted by text length, and, furthermore, a pilot implementation without this length restriction revealed that model latency suffered (>20 seconds per query) when multiple longer example reviews were selected. The range of 75 to 150 words was selected on the basis that these were roughly the lengths of the hand-picked examples used in other prompts and was deemed to be not too long nor too short.

The modified few-shot prompt with N=3 (placeholders where dynamically-retrieved examples would appear) is shown below:

````
Film Review:
```
{dynamically-selected positive review 1}
```
Sentiment: positive

Film Review:
```
{dynamically-selected positive review 2}
```
Sentiment: positive

Film Review:
```
{dynamically-selected positive review 3}
```
Sentiment: positive

Film Review:
```
{dynamically-selected negative review 1}
```
````

```
Sentiment: negative

Film Review:
```
{dynamically-selected negative review 2}
```
Sentiment: negative


Film Review:
```
{dynamically-selected negative review 3}
```
Sentiment: negative


Film Review:
```
{film review text to be classified}
```
Sentiment:
```

b. Traditional Chain-of-Thought

The Chain-of-Thought (CoT) prompt implemented in the original study did not follow the traditional CoT implementation in that, rather than demonstrating a chain of thought that the model should emulate and allowing the model to produce such an output before arriving at the final result, the original study's CoT prompt gave instructions for the logical path that the model should follow, but requested that only the final result be returned. As this means that the SLM would "think silently" and not write out its reasoning first, this variation on CoT prompting might not actually benefit from the chain-of-thought reasoning. Therefore, a more traditional CoT prompt, shown below, was implemented to compare the performance between the two approaches:

```
Q: Is the overall sentiment of the following film review positive or negative?

```
I was really impressed with this film.
The writing was fantastic, and the characters were all rich, and simple.
It's very easy to get emotionally attached to all of them.
The creators of this movie really hit the nail right on the head when it comes to creating
real life characters, and getting the viewer sucked right into their world.
Further, the music is terrific.
They employed some independents to do the score, and some of the soundtrack, and they do a
fantastic job adding to the movie.
If you have a chance to catch this movie in a small theater or at a film festival (like I
did), I highly recommend that you go see it.
Also, on a personal note, Paget Brewster is beautiful in this movie. That's reason enough
to go check it out.
```
A: Let's think step by step. The author of the film review expresses praise of film's
writing, characters, and music.
The author found the characters realistic and the plot engaging, and they recommend seeing
the movie.
The overall sentiment of the review is positive.

Q: Is the overall sentiment of the following film review positive or negative?
```
The beginning was decent; the ending was alright.
Sadly, this movie suffers from complete and utter shallowness of the characters,
unrealistic confrontations/fight scenes, lack of anyone intelligent outside of the shuttle.
This makes for an awful middle screenplay.
Stuff to look for: overly obvious foreshadowing, fast-healing cuts, overly smoky fires, fun
seatbelts, delayed reactions.
```
```

```
I did give it a 4, not a 0, because the start of the movie had some nice elements of
happiness and basic character development.
The relationship between the main, dark-haired girl and her fiancée is touched upon
briefly, and the placement of the blond friend's impact on that relationship is present,
though awkwardly so.
The business discovered at the end is becoming more mainstream and decently done, though,
as another commenter pointed out, not unexpected.  ~viper~
```

```
A: Let's think step by step. Although the author appreciated the start of the film, they
criticized the film's shallow characters, unrealistic scenes, and vapid plot.
The author found the screenplay disappointing and pointed out specific examples of poor
writing and cinematography.
The overall sentiment of the review is negative.


Q: Is the overall sentiment of the following film review positive or negative?
```

```
This film is absolute trash -- in the best way possible. It’s so unapologetically camp that
it borders on iconic.
The writing and acting are hilariously bad; I was laughing so hard I nearly cried.
Honestly, the plot was so predictable and dull that you could zone out completely and still
not miss a thing.
And yet... I’d watch it again in a heartbeat. It’s the perfect movie to throw on with
friends, just to roast every absurd, cringe-worthy moment.
If you love a good "so-bad-it's-good" experience, this is the ultimate guilty pleasure.
```

```
A: Let's think step by step. Even though the author pointed out the film's poor writing,
acting, and predictable plot, they enjoyed the film for its campy, absurd qualities.
The author found the movie entertaining for precisely these flaws and would watch it again
with friends for a laugh.
The overall sentiment of the review is positive.


Q: Is the overall sentiment of the following film review positive or negative?
```

```
{film review to be classified}
```

```
A: Let's think step by step.
```

Rather than explicitly telling the model which thought process to adopt and asking only for the final sentiment label, this more traditional CoT approach shows several examples with corresponding logical analyses of the sentiment. Each CoT example adheres to a set format of “Let’s think step by step” followed by a few salient observations about the reviewer’s experience, culminating with the sentiment label that logically follows from those observations, which primes the SLM to respond in the same way.

Importantly, the prompt demonstrates this analysis as applied to three “flavors” of film reviews. In addition to obvious positive and negative reviews, a third review which could variably be classified as positive or negative, depending on the interpretation, was included. It was observed in initial testing that a number of films were described by the model as containing “mixed” sentiments and that there were many false negatives on reviews in which the author comments on the poor quality of the film but nevertheless enjoys the experience or finds it amusing precisely because the film is so bad. Therefore, a “so bad it’s good” film review example is also included as a guide for how to apply the chain-of-thought reasoning to such reviews – namely, to consider primarily the author’s experience and only secondarily any aesthetic or quality judgments about the film.

Some further “hints” about how to go about analyzing film review sentiment in general were assembled based on an initial inspection of misclassified reviews and added as a primer to the beginning of the prompt, before the series of chain-of-thought examples. Initial testing on a sample from the development set suggested that this improved inference results compared to the CoT prompt without these hints:

```
Consider the following when analyzing film reviews:
- What emotions does the author express toward the film? For example, is the author
impressed, pleased, moved, or excited (positive emotions)? Or is the author bored,
disgusted, disappointed, or confused (negative emotions)?
- Consider the author's description of the film's writing, plot, acting, cinematography,
and other elements. Does the author praise or criticize these aspects? How does this
contribute to the overall sentiment of the review?
- Did the author enjoy the film overall, even if they had some criticisms? This can
indicate a generally positive sentiment. Remember that a film does not need to be high
quality to be enjoyable.
- Consider the author's tone. Does the author use sarcasm or hyperbole for comedic effect?
If so, consider how this affects the overall attitude of the review.
- Would the author recommend the film to others, even if not to everyone? This can indicate
a positive sentiment.
- Does the author mention an explicit star rating out of 10? Ratings >5 are positive and <5
are negative.
```

c. Quantitative Sentiment Analysis

This extended study proposes one additional set of techniques relating to quantitative estimation of film review sentiment. Instead of querying the model to indicate directly whether the review is positive or negative, these prompts query the model to estimate a numeric value associated with the review (e.g. the rating on a scale or number of stars the author might have awarded) and this value is subsequently mapped to the corresponding label as part of postprocessing. Although large and small language models do not typically excel at numeric or mathematical reasoning, it is an interesting experiment to see whether sentiment classification along a numeric scale might be more successful than binary classification as positive or negative, especially given cases of "mixed" sentiment.

These methods combine this quantitative approach with self-consistency prompting, whereby a model is queried multiple times (with `temperature` > 0 to enable varying responses) in order to obtain a sample of responses, and then the majority result or mean of these estimates is taken as the final response. This leverages the variability of the models to achieve an approximation of a majority consensus. In this case, the models were queried three times per film review with `temperature` = 0.4 and the average value was computed. Sentiment was coded as positive if the mean estimated value was greater than 5, otherwise negative.

Three prompts were tested with this quantitative approach:

- **Zero-shot quantitative prompt**

````
Carefully read the following film review and estimate the rating on a scale from 1 to 10
that the author would give to the film or to their overall experience.
Return only the estimated rating as an integer. No further explanation is needed.

```
{review text}
```
````

- **Chain-of-Thought quantitative rating estimation prompt**

The CoT approach to the quantitative prompt merely adapts the newly implemented traditional CoT prompt (with the same examples and hint primer) to the rating estimation task. The model is requested to return its response in the form of a json object containing the chain-of-thought analysis along with the numeric estimation of the review rating. It was hoped that the chain-of-thought reasoning step might help the model arrive at a more suitable numeric estimate as opposed to simply requesting a number. Note that the negative example review explicitly mentions the author's rating, so this same rating (4) was adopted.

```
Q: How would the author of the following review rate the film they watched on a scale from 1 to 10? Return your estimate in json format with "analysis" and "rating" sections.

```
{positive review example}
```
A:
{
  "analysis": "The author of the film review expresses praise of film's writing, characters, and music. The author found the characters realistic and the plot engaging, and they recommend seeing the movie.",
  "rating": 10
}


Q: How would the author of the following review rate the film they watched on a scale from 1 to 10? Return your estimate in json format with "analysis" and "ratings" sections.
```
{negative review example}
```
A:
{
  "analysis": "Although the author appreciated the start of the film, they criticized the film's shallow characters, unrealistic scenes, and vapid plot. The author found the screenplay disappointing and pointed out specific examples of poor writing and cinematography.",
  "rating": 4
}


Q: How would the author of the following review rate the film they watched on a scale from 1 to 10? Return your estimate in json format with "analysis" and "ratings" sections.
```
{“so bad it’s good” review example}
```
A:
{
  "analysis": "Even though the author pointed out the film's poor writing, acting, and predictable plot, they enjoyed the film for its campy, absurd qualities. The author found the movie entertaining for precisely these flaws and would watch it again with friends for a laugh.",
  "rating": 7
}


Q: How would the author of the following review rate the film they watched on a scale from 1 to 10? Return your estimate in json format with "analysis" and "ratings" sections.

```
{review text}
```
A:
```

- **Chain-of-Thought quantitative prompt with rewatch likelihood estimation**

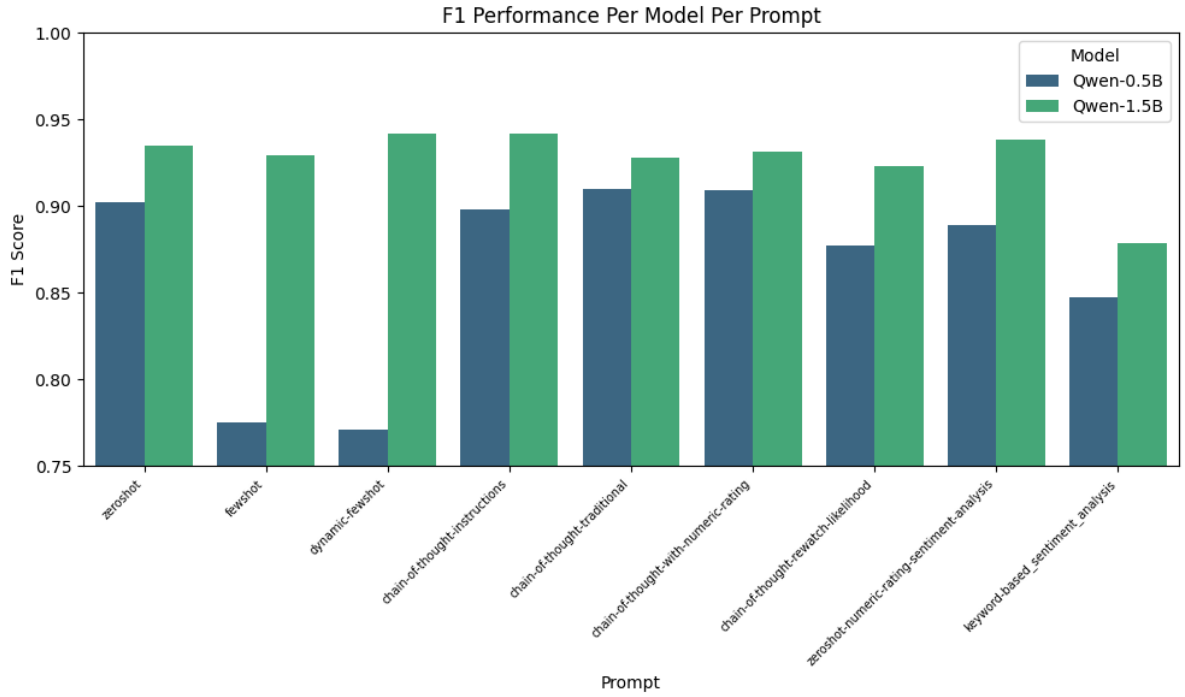
Given the previous discussion of films with “positive” reviews despite explicit mention of poor film quality, and similar mixed cases where, for example, authors express having previously enjoyed a film in the past but being disappointed upon rewatching it, one further quantitative prompting technique was tested. Rather than prompting the model for an estimate of the author’s rating of the film, which might fall into the “so bad it’s good” trap as this focuses on the quality of the film, this prompt requests that the model instead estimate the likelihood that the author would choose to watch the film again in the future, focusing on the reviewer’s overall experience. This seems to be a suitable proxy measure for film review sentiment given that reviewers who are likely to choose to rewatch a film probably also had a positive viewing experience and would have awarded a positive rating, and vice versa. The following

query was used; all other prompt components were identical to the rating estimation CoT prompt, with rating values reinterpreted as likelihood values on the same scale.

```
Q: On a scale from 1 (least likely) to 10 (most likely), how likely is the author to choose
to watch the film again? Return your estimate in json format with "analysis" and
"likelihood to watch again" sections.
```

3. Results

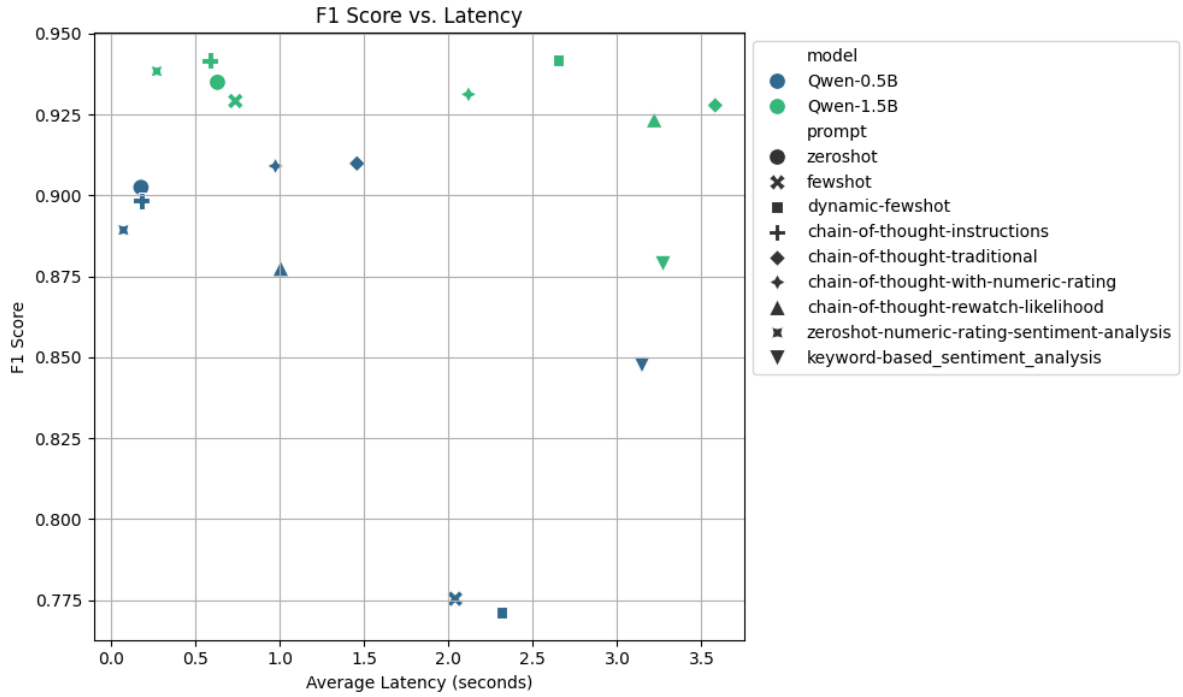
As in the original study, the new prompts were evaluated on a sample of 250 positive and 250 negative reviews from the IMDB dataset test partition, together with the original collection of prompts.[1]



Overall, the top performance is achieved with Qwen2.5-1.5B by the dynamic few-shot and by the instructions-based CoT prompt (from the previous study), which tie for top place at F1 = 0.94. Though they do not secure the top position, all the new prompts from this extended study achieve F1 > 0.92 with Qwen2.5-1.5B. With the smaller Qwen2.5-0.5B model, the traditional CoT prompt introduced in this study achieves the best performance with F1 = 0.91, exceeding the previous top score with this model by the zero-shot prompt from the original study. Oddly, the traditional CoT prompt performs better with the smaller model whereas the instructions-based CoT variation achieves a better score with the larger model.

Perhaps surprisingly given that the dynamic few-shot prompt achieved the best score with the larger SLM, both few-shot prompts scored significantly worse (F1 < 0.8) with the smaller Qwen2.5-0.5B model. Whereas the larger model benefitted from the dynamic similarity-based example retrieval method, this actually worsened the smaller model's performance. The reason for this is not clear, though one guess might be that showing similar examples with opposite labels (as similar positive and similar negative examples are drawn) might confuse the less powerful model and ultimately prove counterproductive.

[1] The original few-shot prompt was slightly restructured in this version to match the format of the currently presented dynamic few-shot prompt, such that the only difference is the choice of examples.



Some further interesting results are the impressive speed and performance of the quantitative self-consistency prompting methods, despite the known limitations of language models' numeric reasoning. Especially the zero-shot version of this approach is very fast. Though an initial test on a smaller development set suggested that estimating the reviewer's likelihood to choose to rewatch the film could be more effective than estimating the rating value directly, this result was not confirmed in the final test results. Still, this quantitative approach to sentiment analysis was overall more effective than the linguistic approach using keyword extraction and analysis from the original study.

4. Conclusion

This SLM sentiment analysis extension has implemented and evaluated five new prompts or prompting methods. The few-shot prompt approach from the original study was enhanced through dynamic retrieval of the most similar film reviews from the train set partition to use as examples, which boosted the few-shot prompt with the Qwen2.5-1.5B model to the top position, tied with the earlier CoT prompt, though this was found to have the opposite effect and worsen performance with the smaller model. The CoT prompt from the earlier study was likewise revisited and reimplemented in a more traditional fashion, which was found to benefit the smaller Qwen2.5-0.5B model, though the original instructions-based CoT prompt variation was more accurate with the larger model. Finally, three prompts were tested with a quantitative approach to sentiment analysis, aiming to estimate the numeric rating of the review or the likelihood that the author would choose to rewatch the film, rather than to directly classify the rating as positive or negative. This approach was combined with self-consistency prompting to leverage model variability to obtain an average estimate, and was found to perform surprisingly well, despite the known limitations on numeric reasoning exhibited by language models.