ICM_CRAG : Comparative Analysis of Approaches

Luka Nedimović
luka.nedimovic@dmi.uns.ac.rs
Faculty of Sciences, University of Novi Sad, Novi Sad
Faculty of Computing, Union University, Belgrade

Abstract

In this short paper, we aim to conduct a comparative analysis of approaches implemented as a part of for the **ICM_CRAG** project. In total, we conduct **30** experiments, on three different datasets, with several hyperparameters.

1 Introduction

This paper aims to explore the effect of different parts of retrieval pipeline, namely effects of chunk size, chunk overlap, maximum number of retrieved chunks and database construction, on three different datasets.

Tokenizer is fixed, as is taken from original implementation: FixedTokenChunker, and the embedding model is fixated to `sentence-transformers/all-MiniLM-L6-v2`. Embedding model can be easily changed through provided bash script.

Datasets we conduct experiments on are the following:

- **Wikitext**. The WikiText language modeling dataset consists of over 100 million tokens from verified Good and Featured articles on Wikipedia.
- **Chatlogs**. The UltraChat 200k dataset is a high quality filtered subset of UltraChat consisting of 1.4M dialogues generated by ChatGPT. Includes all surrounding JSON syntax, making it a more accurate representation of real-world raw text.
- **State of the Union**. Plain transcript of the State of the Union Address in 2024. It is well-structured and clear. This corpus is 10,444 tokens long.

Datasets come with a general `questions_df.csv` evaluation dataset, containing three columns: `question`, `references` (list of relevant chunks for the given question) and `corpus_id` (name of the dataset it belongs to).

The metrics used for evaluation are **Recall** and **Precision**.

2 Experiments

Each experiment comes with a dedicated **bash** script, that can be simply run on-spot, e.g. `./exp_1.sh`. Experiment results are stored in the cmdline-argument-provided `--log` file, which, within the scripts, is set to `icm_rag/experiments/{dataset}/experiments.csv` file.

Experiments store the general experimental information, which contains experiment name, dataset, chunk size, chunk overlap, retriever (database) type, embedding model, k (for top-k retrieval); and experimental results, such as recall and precision (and their standard deviations).

2.1 Experimental Settings

In the following table reside the general experimental settings, for reproduction purposes:

Component	**Details**
GPU	NVIDIA GeForce RTX 3060 (6GB)
CPU	Intel i7-12650H (16) @ 4.600GHz
RAM	16 GB DDR4
Python Version	Python 3.11.11

Table 1: Experimental Settings

2.2 Experiments Results

2.2.1 Wikitext

In the following table, we display the evaluation of our experiments, for the **Wikitext** dataset:

Name	**Ch. Size / Overlap**	**K**	**Recall**	**Precision**
exp_1	800 / 400	10	$\mathbf{97.80} \pm \mathbf{14.33}$	1.13 ± 0.71
exp_2	400 / 120	10	96.16 ± 17.83	2.07 ± 1.27
exp_3	400 / 0	10	95.67 ± 19.03	1.44 ± 0.85
exp_4	300 / 0	10	96.05 ± 16.88	1.90 ± 1.08
exp_5	200 / 0	10	90.46 ± 25.21	$\mathbf{2.64} \pm \mathbf{1.55}$
exp_6	800 / 400	5	$\mathbf{97.80} \pm \mathbf{14.33}$	1.13 ± 0.71
exp_7	400 / 200	5	96.16 ± 17.83	2.07 ± 1.27
exp_8	400 / 0	5	95.67 ± 19.03	1.44 ± 0.85
exp_9	300 / 0	5	96.05 ± 16.88	1.90 ± 1.08
exp_10	200 / 0	5	90.46 ± 25.21	$\mathbf{2.64} \pm \mathbf{1.55}$

Table 2: Experiment Results

We observe the highest recall of $\mathbf{97.80} \pm \mathbf{14.33}$ on experiments $\mathbf{1}$ and $\mathbf{6}$, which both have the same chunk size of $\mathbf{800}$ and overlap of $\mathbf{400}$, whereas the $\mathbf{k}$ is different, which did not impact the answer. These two configurations have stable variance, however, the variance is high in certain cases, e.g. $\pm\mathbf{25.21}$, indicating inconsistency within the system. The highest precision values are achieved in the cases of experiments $\mathbf{5}$ and $\mathbf{10}$, of $\mathbf{2.64} \pm \mathbf{1.55}$, being much higher than the precision values for best-recall models. Admittedly, each of the experiments perform generally well, with mean recall being over $\mathbf{90}\%$.

2.2.2 Chatlogs

In the following table, we display the evaluation of our experiments, for the **Chatlogs** dataset:

Name	**Ch. Size / Overlap**	**K**	**Recall**	**Precision**
exp_1	800 / 400	10	99.38 ± 4.56	1.51 ± 1.21
exp_2	400 / 120	10	$\mathbf{100.00 \pm 0.00}$	2.83 ± 2.23
exp_3	400 / 0	10	99.38 ± 4.56	1.89 ± 1.46
exp_4	300 / 0	10	$\mathbf{100.00 \pm 0.00}$	2.44 ± 1.91
exp_5	200 / 0	10	99.38 ± 4.56	$\mathbf{3.65 \pm 2.82}$
exp_6	800 / 400	5	99.38 ± 4.56	1.51 ± 1.21
exp_7	400 / 200	5	$\mathbf{100.00 \pm 0.00}$	2.83 ± 2.23
exp_8	400 / 0	5	99.38 ± 4.56	1.89 ± 1.46
exp_9	300 / 0	5	$\mathbf{100.00 \pm 0.00}$	2.44 ± 1.91
exp_10	200 / 0	5	99.38 ± 4.56	$\mathbf{3.65 \pm 2.82}$

Table 3: Experiment Results

In this particular set of experiments, we observe certain configurations (namely, experiments $\mathbf{2}$, $\mathbf{4}$, $\mathbf{7}$ and $\mathbf{9}$) having the perfect recall of $\mathbf{100}\%$. Other experients still achieve a high recall, of $\mathbf{99.38}\%$, with small variance of ± 4.56. Experiments $\mathbf{5}$ and $\mathbf{10}$ have the highest precision of 3.65 ± 2.82, indicating that these experiments retrieved several relevant chunks per query, but with a high degree of variability.

2.2.3 State of the Union

In the following table, we display the evaluation of our experiments, for the **State of the Union** dataset:

Name	**Ch. Size / Overlap**	**K**	**Recall**	**Precision**
exp_1	800 / 400	10	94.05 ± 22.96	0.65 ± 0.46
exp_2	400 / 120	10	$\mathbf{100.00 \pm 0.00}$	1.32 ± 0.86
exp_3	400 / 0	10	93.44 ± 23.41	0.97 ± 0.69
exp_4	300 / 0	10	93.35 ± 23.53	1.29 ± 0.91
exp_5	200 / 0	10	98.04 ± 11.98	$\mathbf{2.03 \pm 1.30}$
exp_6	800 / 400	5	94.05 ± 22.96	0.65 ± 0.46
exp_7	400 / 200	5	$\mathbf{100.00 \pm 0.00}$	1.32 ± 0.86
exp_8	400 / 0	5	93.44 ± 23.41	0.97 ± 0.69
exp_9	300 / 0	5	93.35 ± 23.53	1.29 ± 0.91
exp_10	200 / 0	5	98.04 ± 11.98	$\mathbf{2.03 \pm 1.30}$

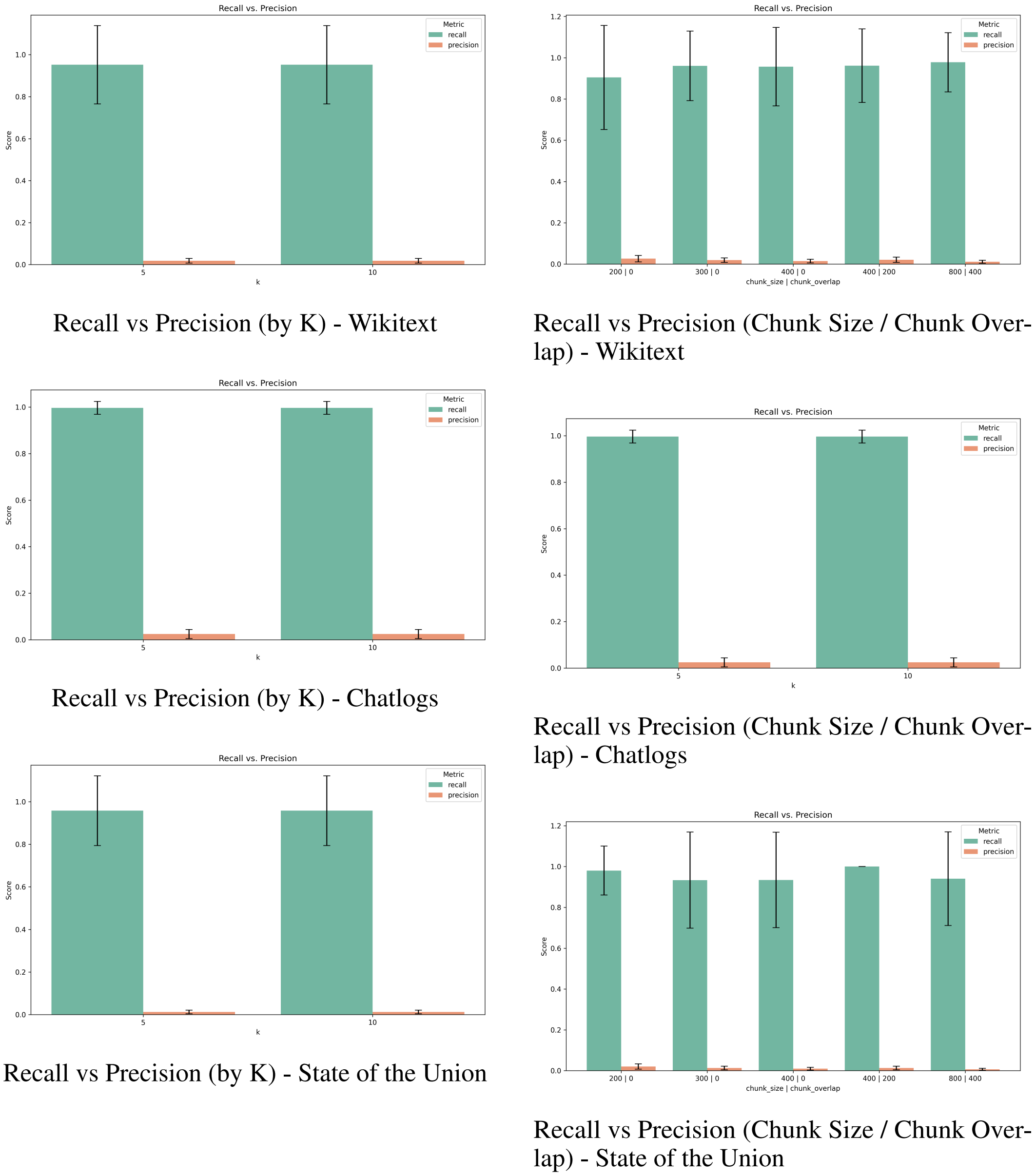
Table 4: Experiment Results

Experiments $\mathbf{2}$ and $\mathbf{7}$ achieve the perfect recall of $\mathbf{100}\%$, whilst the other configurations achieve $> \mathbf{93}\%$. Howeever, most of the configuartions have variance higher than $\mathbf{20}\%$, indicating possbile inconsistent behavior. The highest scoring experiments do not have the maximal chunk size (i.e. $\mathbf{800/400}$). We observe lowest recall with chunk size of $\mathbf{300}$, and no overlap. Precision-wise, experiments $\mathbf{5}$ and $\mathbf{10}$ score the highest on precision, whilst the other model are generally about at minimum $\mathbf{0.74}\%$ away.

3 Conclusion

In this short paper, we conduct $\mathbf{30}$ experiments on file retrieval task. Experimental evidence suggests that medium sized chunks work better for datasets such as Chatlogs and State of the Union, whilst the bigger size chunks work better for Wikitexts. Precision is decreasing by increasing the size of the chunk, and even chunks of very small sizes (such as $\mathbf{200}$) can lead to high performance of $\mathbf{98}\%$.

4 Appendix



Recall vs Precision (by K) - Wikitext

Recall vs Precision (Chunk Size / Chunk Overlap) - Wikitext

Recall vs Precision (by K) - Chatlogs

Recall vs Precision (Chunk Size / Chunk Overlap) - Chatlogs

Recall vs Precision (by K) - State of the Union

Recall vs Precision (Chunk Size / Chunk Overlap) - State of the Union

Figure 1: Additional plots showing Recall and Precision under different factors