

# Supplementary Materials: Coarse-to-Fine Retrieval and Multi-modal Efficient Tuning for Document VQA(CREAM)

Anonymous Author(s)

## A ADDITIONAL ANALYSIS ON PERFORMANCE

### A.1 Text Embeddings of CREAM

The success of our retrieval process will depend to some extent on the quality of the text embeddings, emphasizing the need for an embedding model with superior semantic matching capabilities. In order to highlight the performance of the text embedding, we choose the text chunk of 512 tokens to conduct experiments and compare it to three open-source embedding models, including e5-large, instructor-large, and bge-large. As indicated in Figure 1, The comparative evaluation revealed that bge-large consistently outperforms the others in terms of average performance. Therefore, we ultimately selected bge-large as the retrieval model for all datasets.

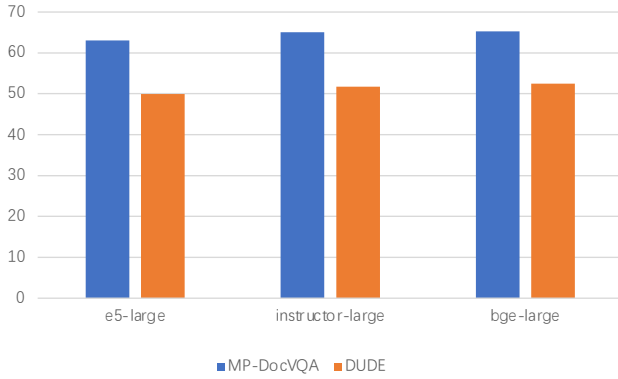


Figure 1: The effect of different text embedding models

### A.2 LLM Ranking strategies of CREAM

Given that the pre-trained model we utilized, RankVicuna [1], is limited to processing only 8 text chunks of 512 tokens each at a time, and since most multi-page documents contain more than 8 text chunks, we addressed this limitation through grouping. In the experiment, to validate the effectiveness of our method, we compared different reranking strategies employed by the LLM. As demonstrated in Table 1, the text chunks selected after multiple rounds of ranking show more effective results than those chosen following a single round of ranking.

### A.3 Further Analysis of CREAM

Similar to word embedding, different page embedding relationships can influence semantic interpretation. Therefore, we assess the performance of the page embedding module as shown in Table 2. Simultaneously, after conducting retrieval, we select the three most relevant pages due to resource and model acceptance length

Table 1: Comparison of ranking strategies in LLM

Dataset	Single-round	Multi-round
MPDocVQA	63.72	65.32
DUDE	51.34	52.46

Table 2: Effect of page embedding and number of pages

page Embedding	MPDocVQA			DUDE		
	1	2	3	1	2	3
✓	64.46	64.84	65.32	51.55	52.21	52.46
✗	64.01	64.42	64.28	50.88	51.43	51.79

constraints. The continuity of the relevant pages confirms their pertinence to the query. The experimental results reveal that page embedding has a impact when dealing with several pages. Using just one page results in lower performance, whereas the difference in results between two and three pages is minimal, suggesting that the content of two or three pages generally contains the information pertinent to the query.

### A.4 Compared with more methods

Table 3 enumerates the current state-of-the-art models in terms of performance, and the results indicate that our approach maintains its status as the leading state-of-the-art on all datasets, except for DocVQA. Upon analyzing the error instances in the DocVQA dataset, it was observed that most errors arise from the failure to recognize the layout and visual elements of the document, which is a known limitation of OCR and the vision encoder. Moving forward, we aim to delve deeper into the multi-modal aspects of the framework to address and mitigate this issue.

## B MORE QUALITATIVE RESULTS

### B.1 Results from different datasets

CREAM demonstrates the ability to accurately identify corresponding values in charts with complex layouts, which likely reflects the inherent robustness of the document vision encoder and the large language model. As illustrated in Figure 3 a, The model effectively comprehends a document page featuring multiple charts. Although the sequences identified by OCR lacked document layout information, our implementation of a visual encoder equipped with document image understanding capabilities played a complementary role.

Concurrently, it is noteworthy that despite the expansion in the acceptable token length for LLMs, certain limitations persist. These models often exhibit sensitivity to the beginning and end of sentences and may selectively overlook some contextual information

Table 3: Performance comparison between CREAM and state-of-the-art methods.

DocVQA	InfoVQA	VisualMRC	MP-DocVQA	DUDE
87.8(DocFormerv2)	48.8(DocFormerv2)	364.2(LayoutT5)	62.0(HiVT5)	50.0(DocGptVQA)
79.4	53.6	377.9	65.3	52.5

in the middle. Therefore, the retrieval of pertinent information and the condensation of context remain critical steps in the process. As shown in Figure 3 b, CREAM first locates the context related to the question, thereby improving the accuracy of the answer.

B.2 Examples of different text embedding

Figure 4 and Figure 5 illustrate the distinct performances of three different embedding models. Here we set up a text embedding only retrieval with a text chunk length of 512 tokens. It is observed that bge-large demonstrates the highest accuracy in context localization, followed by instructor-large, with e5-large ranking last. The efficacy of the embedding model is crucial in the retrieval module, showing

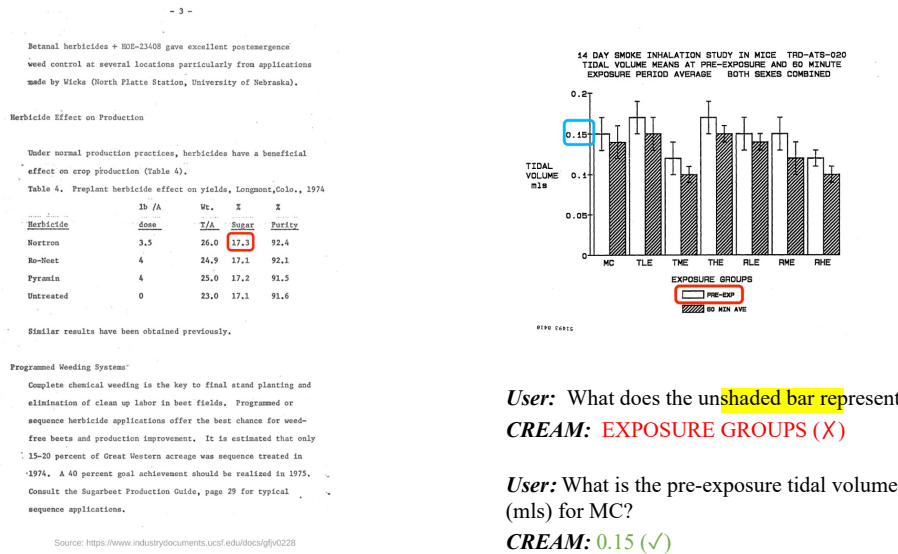
a positive correlation with the overall performance in document-based question answering.

B.3 Examples of different retrieval methods

Figure 6 and Figure 7 present a comparison between the outcomes of our coarse-to-fine retrieval and text embedding only retrieval. The examples clearly indicate that the content retrieved through our method exhibits greater accuracy. By applying a filter through LLM, we can obtain a context of the higher quality.

REFERENCES

[1] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *arXiv preprint arXiv: 2309.15088* (2023).



User: What does the unshaded bar represent?

CREAM: EXPOSURE GROUPS (X)

User: What is the pre-exposure tidal volume (mls) for MC?

CREAM: 0.15 (✓)

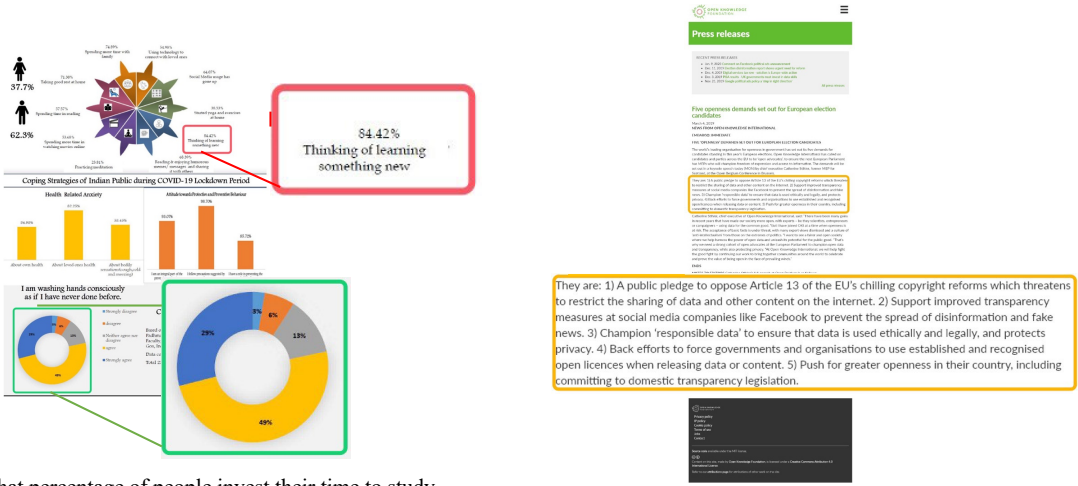
User: what is the % of sugar in norton?

CREAM: 17.3 (✓)

(a)

(b)

Figure 2: In the case of the DocVQA dataset, the section encircled in the box represents the answer.



User: What percentage of people invest their time to study something new?

CREAM: 84.42% (✓)

User: What percentage of people agree with washing their hands often ever than before?

CREAM: 49 (✓)

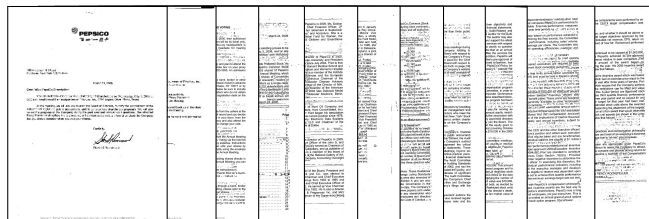
(a)

User: What is the third demand in Catherine Stihler's keynote speech?

CREAM: The third demand is to champion responsible data to ensure that data is used ethically and legally, and protects privacy. (✓)

(b)

Figure 3: In the case of the InfoVQA and VisualMRC datasets, Crucial regions are enlarged for clearer visualization.

**Bge-large:**

**Context retrieved:** by the Please vote your shares promptly through any of the means described on the proxy card. Board for election at ..... The Governance Committee shall review the desired experience, account the current Board will generally look for individuals who have displayed high ethical Board members and the specific needs of the Corporation and the standards, Larry D. Thompson Secretary A-1 26 integrity, and ..... best interests under of the Enrollment " links and follow the instructions provided. to circumstances and (3) in a manner he or she believes to be .....

**Question:** who is the secretary of the company?

**Answer:** Larry D. Thompson

**Real answer:** Larry D. Thompson

**instructor-large:**

**Context retrieved:** Our business and affairs are overseen by our 34,869 shares; James J. Schiro, 10,377 shares; Franklin A. Thomas, 39,244 shares; Cynthia Business Corporation- Act and our By-Laws. Members of the Board of Directors are kept informed of the M. Trudell, 26,866 shares; Daniel Vasella, 10,417 shares; Michael D. White ..... Prior various engineering and manufacturing supervisory positions. In 1995, Co. as' chemical process engineer. In 1981, she joined GM and held she to joining PepsiCo, Ms. Nooyi spent four years as Senior Vice President of Strategy, Planning and .....

**Question:** who is the secretary of the company?

**Answer:** M. Trudell

**Real answer:** Larry D. Thompson

**e5-large-v2:**

**Context retrieved:** to solicit proxies from their principals and will pay the brokers and bank brokers Oversee and interact, with senior, management with respect to key aspects succession, of the nominees their expenses for the solicitation. Employees of PepsiCo may also solicit proxies. They will business 'including strategic planning, management development and not receive any additional pay for the solicitation .....

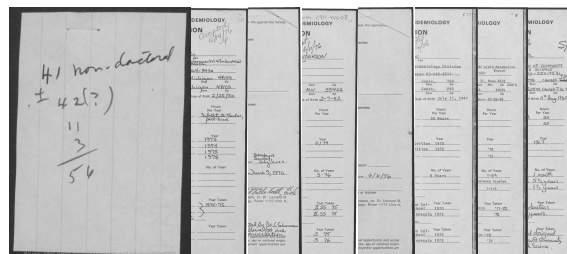
It is Board policy that executive officers and other members of senior management who report directly to the CEO be present at Board meetings at the invitation of the .....

**Question:** who is the secretary of the company?

**Answer:** William J. Darby

**Real answer:** Larry D. Thompson

Figure 4: Case comparison of different embedding models. The bge-large is superior to instructor-large and e5-large.

**Bge-large:**

**Context retrieved:** and Abnormalities M-W-F 17. Please make the following room reservation for me: / Middlebrook Hall (Two room, twin beds) Arrive 6/20 Depart 7/10 in a 18. Signature of Applicant Mah D. Andow Date 3/25 19. I approve of this application Pr.s. Pasternock Department Chairman or Adviser Send this form with check for \$25.00, made payable to the University of Minnesota, to: Dr. Leonard M. Schuman ..... Harmard Course Title School in Public Hours Health School Year Taken 15. Have you had any graduate level courses in epidemiology?

**Question:** What is the arrival date?

**Answer:** 6/20

**Real answer:** 6/20 or 6-20

**instructor-large:**

**Context retrieved:** Please make the following room reservation for me: XX Middlebrook Hall (Two in a room, twin beds) Arrive 6-20 Depart 7-10 18. Signature of Applicant Lusly SRisbord Date 4/25/26 19. I approve of this application Cary H. Spincy Department (Assist prf Chairman Adviser) Adviser or Send this form with check for \$25.00, made payable to the University of Minnesota, to: Dr. Leonard M. Schuman, Program Director, Epidemiology Summer Session, Division of Epidemiology, Room 1-117 Unit A, .....

**Question:** What is the arrival date?

**Answer:** 6-20

**Real answer:** 6/20 or 6-20

**e5-large-v2:**

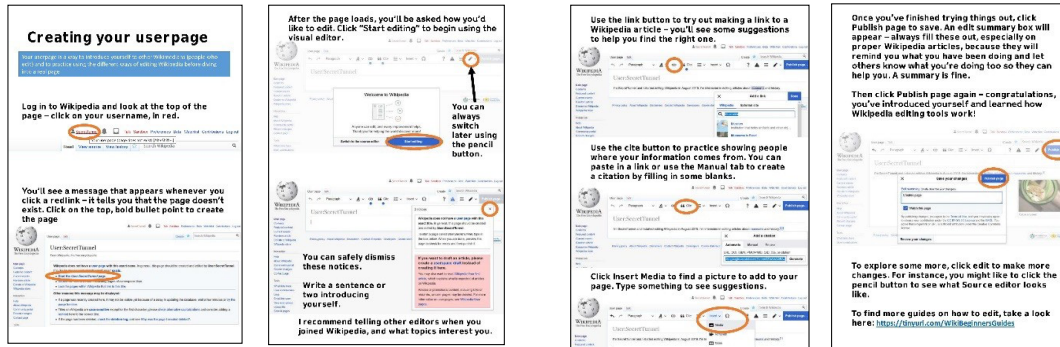
**Context retrieved:** Office Address No. Street City State Zip 7. Home Address 323 4th ST S.E. MNCLS MIN 55414 No. Street City State Zip 8. Home Phone 3381936 9. Sex Female 10. Date of Birth 4/2152 11. Teaching Activities Types of Hours School or Agency Subject Students Per Year 12. College, University, and Professional Education School Major Degree Year LALUPENCE UNIV BIOLOGY BA 1974 13. Professional Experience (Start with Present Position) Employer Position Duty No. of Years UNIU OF wis. LAB T ECH ANTIBIOTIC RESEARCH 2 14. ....

**Question:** What is the arrival date?

**Answer:** 7/27/76

**Real answer:** 6/20 or 6-20

Figure 5: Case comparison of different embedding models. The bge-large is superior to instructor-large and e5-large. Bge-large and instructor-large are better than e5-large.



### Text Embedding+RankLLM

**Context retrieved:** editing articles about museums and history. Featured content X Add a citation Current events Random article Automatic Manual Donate to Wikipedia Privacy policy About Wikipedia Disclaimers Contact Wikipedia Developers Cookie statement Reuse Wikipedia store URL, DOI, ISBN, PMC/PMID, QID, title, or citation Interaction .....Generate Help Click [Insert Media](#) to find a picture to add to your page. Type something to see suggestions. SecretTunnel Talk Sandbox Preferences Beta Watchlist Contributions Log out User page Talk .....

**Question :** What is the key to find a picture?

**prediction :** insert media

**Real answer:** Insert Media

### Text Embedding Only

**Context retrieved:** Review your changes contributions To explore some more, click edit to make more changes. For instance, you might like to click the pencil button to see what Source editor looks like..... in editing articles about museums and history Featured content Add a link Done Current events Random article Donate to Wikipedia Privacy policy About Wikipedia Disclaimers Contact Wikipedia Developers Cookie statement Wikipedia External site Wikipedia store Museums Interaction Help Museum About Wikipedia Institution that holds artifacts and other objects .....

**Question :** What is the key to find a picture?

**prediction :** the magnifying glass icon

**Real answer:** Insert Media

Figure 6: Case comparison of different retrieval strategies. The text embedding only retrieval locates an error message.



### Text Embedding+RankLLM:

**Context retrieved:** J. Balbus, J.L. Gamble, C.B. Beard, J.E. Bell, D. .... PHOTO CREDITS port-on-the-impacts-of-climate-change-on-human-health-in cover and title page-Manhattan skyline: iStockPhoto.com/ the-united-states stockelements; Farmer: Masterfile/Corbis; Girl getting checkup: Rob Lewine/Tetra Images/Corbis 3. 2014: Climate Change Impacts in the [United States](#): The Third Pg. vii-Elderly Navajo woman and her niece .....

Ch.9: Human health. Climate Change Impacts in the [United States](#): The Third National Climate Assessment. Melillo, J.M., T.!,

**Question :** which country specified in this document?

**Prediction :** united states

**Real answer:** United States

### Text Embedding Only:

**Context retrieved:** Centers for Disease Control and Prevention Rupa Basu, California Office of Environmental Health Hazard and Ross Bowling, Office of the Assistant Secretary for Administration Assessment Kathleen Danskin, Office of the Assistant Secretary for Preparedness Paul English, Public Health Institute, Oakland, CA and Response Kim Knowlton, Columbia University Mailman School of Public Health Stacey Degra...

medium consensus Unlikely Medium Two kinds of language are used when describing the 1 in 3 Suggestive evidence ..... Office of Air and Radiation Michelle Hawkins, National Oceanic and Atmospheric Administration.

**Question :** which country specified in this document?

**Prediction :** none

**Real answer:** United States

Figure 7: Case comparison of different retrieval strategies. The text embedding only retrieval could not locate the relevant information.