Supplemental Materials

1: Data Location

2: Author Statement

3: Hosting Plan

4: DOIs

5: Datasheet for CRAG

1 Dataset Locations

CRAG is available at `https://github.com/facebookresearch/CRAG/`.

2 Author Statement

We bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

3 Hosting, Licensing, and Maintenance Plan

This project is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). This license permits sharing and adapting the work provided it is not used for commercial purposes and appropriate credit is given. We host the dataset at `https://github.com/facebookresearch/CRAG/` and commit to maintaining CRAG to serve research communities in advancing RAG solutions and general QA solutions.

4 DOIs

Please refer to Section 3 for our hosting plan. Hosting our datasets and code in our institutional repository provides sufficient accessibility, persistence, and traceability, which sufficiently addresses the needs that a DOI typically serves.

Institutional Support and Stability: Our repository is supported by our institution, which ensures its long-term stability and reliability. We commit to maintaining the repository and its contents, making it a dependable resource for accessing our datasets and code.

Version Control and Update Mechanisms: Our repository makes it easier for us to ensure transparency and reproducibility of our work, by allowing users to easily access not only the most current version of the data and code but also previous versions and update history.

Metadata and Documentation: The dataset and code in our repository are accompanied by detailed metadata and documentation. This documentation includes all necessary information to understand and use the data and code effectively and ensures that users can properly attribute the resources used in their research, fulfilling a key role of the DOI.

5 Datasheet for CRAG

In this section, we use the framework of Datasheets for Datasets [**?**] to form a datasheet for CRAG, aiming to document the motivation, composition, collection process, recommended uses, and other information for our CRAG benchmark.

5.1 Motivation

Q1. **For what purpose was the dataset created? Was there a specific task in mind?**

To provide a robust and comprehensive evaluation benchmark for testing RAG systems.

Q2. **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The authors of the paper.

Q3. **Who funded the creation of the dataset?**

Meta Platforms.

Q4. **Any other comments?**

No.

5.2 Composition

Q5. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

Each instance corresponds to a factual question-answering pair.

Q6. **How many instances are there in total (of each type, if appropriate)?**

CRAG contains 4,409 question-answering pairs.

Q7. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

The CRAG questions are constructed based on 1) human written templates and entities sampled from knowledge graphs(KGs); 2) human written questions which likely have an answer from web search.

Q8. **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?**

Each instance contains a natural language question and a corresponding answer.

Q9. **Is there a label or target associated with each instance?**

Yes. Please refer to Section **??**.

Q10. **Is any information missing from individual instances?**

No.

Q11. **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

Yes.

Q12. **Are there recommended data splits (e.g., training, development/validation, testing)?**

Yes. Please refer to Appendix **??**.

Q13. **Are there any errors, sources of noise, or redundancies in the dataset?**

We conducted multiple rounds of human review and strived to make CRAG high quality. In case there were minor errors existing in the question or answer, we will fix them in future releases.

Q14. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

CRAG is self-contained, and does not reply on external resources to use.

Q15. **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?**

No.

Q16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

No.

Q17. **Does the dataset relate to people?**

No.

Q18. **Does the dataset identify any subpopulations (e.g., by age, gender)?**

No.

Q19. **Is it possible to identify one or more natural persons, either directly or indirectly (i.e., in combination with other data) from the dataset?**

No.

Q20. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

No.

Q21. **Any other comments?**

No.

5.3 Collection Process

Q22. **How was the data associated with each instance acquired?**

Please refer to Section **??**.

Q23. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

Please refer to Section **??**.

Q24. **If the dataset is a sample from a larger set, what was the sampling strategy?**

Part of the CRAG data were created based on KGs, the entities were sampled randomly from the head, torso and tail popularity groups in the KGs. Please refer to Section **??** for more details.

Q25. **Who was involved in data collection process (e.g., students, crowd-workers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Students, linguists, and engineers are involved in the data collection process.

Q26. **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**

Feb to Apr 2024. Yes, the timeframe matches the creation timeframe of the data associated with the instances.

Q27. **Were any ethical review processes conducted (e.g., by an institutional review board)?**

Not applicable. CRAG does not contain sensitive information related to humans, as a result, an ethical review is not needed.

Q28. **Does the dataset relate to people?**

No.

Q29. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

Not applicable.

Q30. **Were the individuals in question notified about the data collection?**

Not applicable.

Q31. **Did the individuals in question consent to the collection and use of their data?**

Not applicable.

Q32. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

Not applicable.

Q33. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

Not applicable.

Q34. **Any other comments?**

No.

5.4 Preprocessing, Cleaning, and/or Labeling

Q35. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

Yes, we conducted multiple review and cleaning process to ensure the data quality.

Q36. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

Yes.

Q37. **Is the software used to preprocess/clean/label the instances available?**

Yes.

Q38. **Any other comments?**

No.

5.5 Uses

Q39. **Has the dataset been used for any tasks already?**

Yes, the CRAG benchmark has been used in the 2024 Meta KDD Cup Challenge.

Q40. **Is there a repository that links to any or all papers or systems that use the dataset?**

Yes, please refer to "Meta KDD Cup 2024 Overview" in Section 1 for more details.

Q41. **What (other) tasks could the dataset be used for?**

Any RAG or QA evaluation task.

Q42. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

No.

Q43. **Are there any tasks for which the dataset should not be used?**

No.

Q44. **Any other comments?**

No.

5.6 Distribution

Q45. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

Yes, we have released the CRAG dataset.

Q46. **How will the dataset be distributed (e.g., tarball on website, API, GitHub)**

CRAG is available at `https://github.com/facebookresearch/CRAG/`.

Q47. **When will the dataset be distributed?.**

We have already released the valication and public test set for the 2024 Meta KDD Cup challenge.

Q48. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

CRAG is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). Please refer to Section 3 for more details.

Q49. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

No.

Q50. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

No.

Q51. **Any other comments?**

No.

5.7 Maintenance

Q52. **Who will be supporting/hosting/maintaining the dataset?**

The authors of the paper.

Q53. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

By emailing to the correspondence author.

Q54. **Is there an erratum?**

No.

Q55. **Will the dataset *be updated (e.g., to correct labeling errors, add new instances, delete instances)?* If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?**

We maintain the dataset in the GitHub data repository.

Q56. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

Not applicable.

Q57. **Will older versions of the dataset continue to be supported/hosted/maintained?**

No.

Q58. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

Not currently.

Q59. **Any other comments?**

No.