Successful Case	Failure Case
Context: (QReCC Session 17) q_1: What are the different forms of energy? r_1: Examples of these are: light energy, heat energy, mechanical energy, gravitational energy, electrical energy, sound energy, chemical energy, nuclear or atomic energy and so on. q_2: How can it be stored? r_2: Batteries, gasoline, natural gas, food, water towers, a wound up alarm clock, a Thermos flask with hot water and even pooh are all stores of energy. They can be transferred into other kinds of energy. q_3: What type of energy is used in motion? r_3: Motion energy – also known as mechanical energy – is the energy stored in moving objects. As the object moves faster, more energy is stored. q_4: Tell me about mechanical energy. r_4: Mechanical energy is the sum of kinetic and potential energy in an object that is used to do work. In other words, it is energy in an object due to its motion or position, or both. **Current Query**: q_5: Give me some examples. **Human-Rewritten**: q_5^*: Give me some examples of mechanical energy. **ConvGQR Reformulated Query**: $\hat{q_5}$: Give me some examples of mechanical energy. The energy in a motor is the sum of kinetic and potential energy in an object that is used to do work. **Relevant Passage**: p^*: Objects have mechanical energy if they are in motion ... *A few examples are: a moving car possesses mechanical energy due to its motion(kinetic energy) and a barbell ... its vertical position above the ground(potential energy).* **Dense Score**: 0.33 (Human-Rewritten) **1.00 (Ours)** **Sparse Score**: 0.03 (Human-Rewritten) **0.17 (Ours)**	**Context**: (QReCC Session 5) q_1: What are the best ways to cook a turkey? r_1: Heat the oven to 450°F to preheat and then drop the temperature to 350°F when putting the turkey into the oven. The turkey is done when it registers a minimum of 165° in the thickest part of the thigh. q_2: Should I brine a turkey before smoking it? r_2: Use a brine before smoking to help keep meat moist while cooking and to add flavor. **Current Query**: q_3: How much salt do I use to brine it? **Human Oracle Rewrite**: q_3^*: How much salt do I use to brine a turkey? **ConvGQR Reformulated Query**: $\hat{q_3}$: How much salt do I use to brine a turkey? Salt: 1 teaspoon per pound of turkey breast, 1 teaspoon per pound of ground turkey breast, 1 teaspoon per pound of ground turkey breast, **Relevant Passage**: p^*: How To Brine a Turkey ... Salt Solution *The basic ratio for turkey brine is two cups of kosher salt to two gallons of water. Some recipes include sweeteners or acidic ingredients to balance the saltiness.* **Dense Score**: **1.00** (Human-Rewritten) 0.5 (Ours) **Sparse Score**: 0.13 (Human-Rewritten) 0.00 (Ours)

Table 8: Two additional concrete examples about different effectiveness of expanding generated query. The blue tokens and the orange tokens stand for the rewritten query and the expanded query of ConvGQR. The expansion terms and the gold answer are underline and *italicized* in the relevant passage.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 6 (Limitation)

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4 (Experiments)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.