3.1 Generate step-by-step explanation

For each question, the LLM is required to generate a definitive answer followed by a step-by-step explanation. We perform the experiment on two types of questions: those that require a ternary response (affirmative, negative, or uncertain) and those that present multiple-choice options. The definitive answer will be in the form of ”yes,” ”maybe,” or ”no” for the first type of questions, or a selection from the multiple-choice options for the second type. This is followed by generating a detailed step-by-step explanation for the chosen answer, which is critical for the subsequent verification chain. The step-by-step breakdown converts the model’s reasoning into discrete units that can be independently verified for truthfulness and consistency, thereby enabling an estimation of the overall confidence in the response.

3.2 Plan verification

Upon generating the initial answer and step-by-step explanation, the model proceeds to formulate a set of verification questions, with each one targeting a single step in the explanation. These questions are purposefully designed to challenge the accuracy of particular factual claims within the individual steps of the explanation. The objective of these questions is to verify the truthfulness of each assertion without necessitating supplementary knowledge or additional context for their resolution. For example, in response to the statement *”Ringed sideroblasts are a characteristic feature of iron overload, particularly in the bone marrow”*, a potential verification question could be *“What condition are ringed sideroblasts typically indicative of?”* This question directly targets the factual claim made within the statement and is structured to elicit a response that either confirms or refutes the accuracy of the original statement.

While the model is capable of formulating reasonable verification questions on a zero-shot instruction, the incorporation of a few-shot prompt significantly refines this procedure by enhancing the efficacy of the verification questions. A few-shot prompt presents the model with a set of carefully curated exemplary pairs of statements and corresponding verification questions. These examples serve as a template, showcasing the structure and purpose of a well-crafted verification question. Consequently, this few-shot prompt empowers the model to formulate questions that are not only relevant but also incisive in their ability to discern and test the validity of factual assertions.

3.3 Execute verification

Given the verification questions, in the next step, the model executes the verification procedure to self-check whether the explanation is accurate. We examined several different approaches for verification in our experiment.

3.3.1 Step verification

As a base for the verification procedure, we use the model to directly assess the truthfulness and the consistency related to the previous steps of each sentence in the explanation without utilizing the verification questions. For each sentence, the model is prompted to determine its truthfulness based on the prior sentences in the explanation, classifying it as true or false. This serves as a baseline measurement of the model’s ability to self-validate its content.

This direct approach assumes that the language model is intrinsically capable of recognizing factual information. It provides a straightforward validation mechanism without the additional layer of complexity introduced by verification questions.

3.3.2 CoVe

In this approach, the LLM answers the verification questions independently to avoid the influence of the initial output. Next, the independent answer will be checked against the original statement being examined for consistency. This is performed by providing the model with both the answer and the statement and asking it to decide if they are consistent or not.

The assumption for this approach is that the model is less likely to repeat any hallucinations present in the initial explanation when answering the verification questions independently without any context. If the independent response aligns with the explanation, the corresponding statement has a lower possibility of being a hallucination. On the contrary, an inconsistency between the two indicates a potential error or hallucination in the explanation, making the initial answer less plausible.

3.3.3 Two-phase verification

In this more sophisticated approach, the model is prompted to answer each verification question twice. First, the model answers the verification question independently, as in the previous approach. Next, the model is given the statement to be verified as the context and prompted to answer the verification question again. To evaluate whether the two answers are consistent, we adopt a method for checking semantic equivalency which uses a Deberta-large model (He et al., 2021) for a bidirectional entailment check (Kuhn et al., 2023). This process involves appending a special token between the answers and evaluating whether each answer can be inferred from the other, with equivalence determined by mutual ”entailment” classifications by the model. An example of the Two-phase Verification procedure is illustrated in Figure 2.

The rationale for integrating the second verification question answering step responds to two significant challenges encountered during the consistency check in CoVe:

1. Ambiguity in Consistency Checks: The instructions for a consistency check can themselves be ambiguous due to the different interpretations of “consistency”. Additionally, the model may fixate on superficial linguistic patterns rather than the underlying factual content. Various phrasings conveying the same meaning may not be recognized as consistent by the model, leading to false judgments in identifying consistency.
2. Relevance and Information Discrepancies: The independent answer generated by the model could introduce additional information that is not strictly relevant to the initial explanation, or it could omit crucial details, making it difficult to accurately assess consistency. The answer might be factually correct in itself but still not align perfectly with the explanation due to differences in scope or detail level.

3.4 Uncertainty quantification

Following the completion of the verification steps, we translate the findings into a measurable indicator of uncertainty. This involves counting the statements identified as inconsistent in the verification phase, relative to the overall number of statements in the provided explanation. To express this quantitatively, we compute the Uncertainty Level (UL) using the formula below:

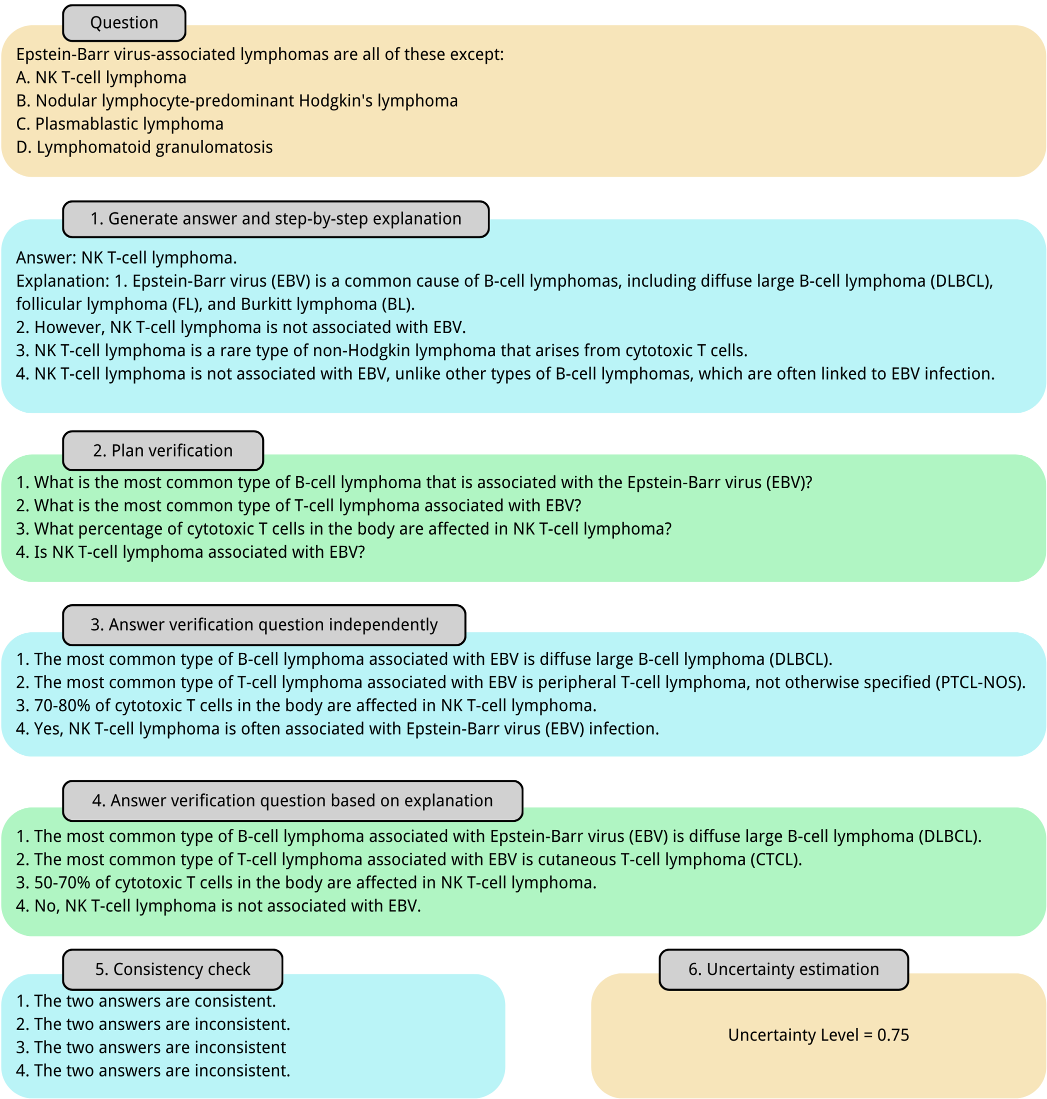
$$UL = \frac{\text{Number of Inconsistent Statements}}{\text{Total Number of Statements in Explanation}}$$

4 Experiment

4.1 Experimental setting

Models We conduct the experiment on Llama 2 Chat (Touvron et al., 2023), which is a collection of open-source chat models fine-tuned for optimized dialogue use cases. Llama 2 Chat (7b) and Llama 2 Chat (13b) were examined in our experiment.

Datasets We consider three biomedical question-answering (QA) datasets: PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). **PubMedQA** is a biomedical research QA dataset designed to answer questions with a yes/no/maybe format. Each question comes with a context extracted from the corresponding abstract of a research paper and challenges models to reason over quantitative biomedical content. The expert-annotated questions were utilized in our experiment. **MedQA** is a free-form multiple-choice



Question

Epstein-Barr virus-associated lymphomas are all of these except:
A. NK T-cell lymphoma
B. Nodular lymphocyte-predominant Hodgkin's lymphoma
C. Plasmablastic lymphoma
D. Lymphomatoid granulomatosis

1. Generate answer and step-by-step explanation

Answer: NK T-cell lymphoma.
Explanation: 1. Epstein-Barr virus (EBV) is a common cause of B-cell lymphomas, including diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), and Burkitt lymphoma (BL).
2. However, NK T-cell lymphoma is not associated with EBV.
3. NK T-cell lymphoma is a rare type of non-Hodgkin lymphoma that arises from cytotoxic T cells.
4. NK T-cell lymphoma is not associated with EBV, unlike other types of B-cell lymphomas, which are often linked to EBV infection.

2. Plan verification

1. What is the most common type of B-cell lymphoma that is associated with the Epstein-Barr virus (EBV)?
2. What is the most common type of T-cell lymphoma associated with EBV?
3. What percentage of cytotoxic T cells in the body are affected in NK T-cell lymphoma?
4. Is NK T-cell lymphoma associated with EBV?

3. Answer verification question independently

1. The most common type of B-cell lymphoma associated with EBV is diffuse large B-cell lymphoma (DLBCL).
2. The most common type of T-cell lymphoma associated with EBV is peripheral T-cell lymphoma, not otherwise specified (PTCL-NOS).
3. 70-80% of cytotoxic T cells in the body are affected in NK T-cell lymphoma.
4. Yes, NK T-cell lymphoma is often associated with Epstein-Barr virus (EBV) infection.

4. Answer verification question based on explanation

1. The most common type of B-cell lymphoma associated with Epstein-Barr virus (EBV) is diffuse large B-cell lymphoma (DLBCL).
2. The most common type of T-cell lymphoma associated with EBV is cutaneous T-cell lymphoma (CTCL).
3. 50-70% of cytotoxic T cells in the body are affected in NK T-cell lymphoma.
4. No, NK T-cell lymphoma is not associated with EBV.

5. Consistency check

1. The two answers are consistent.
2. The two answers are inconsistent.
3. The two answers are inconsistent
4. The two answers are inconsistent.

6. Uncertainty estimation

Uncertainty Level = 0.75

Figure 2: Illustration of Two-phase Verification process with an example question

QA dataset collected from questions in professional medical board examinations, such as the United States Medical Licensing Examination (USMLE). **MedMCQA** is a large-scale multiple-choice QA dataset derived from real-world medical entrance exam questions. It is designed to test a variety of reasoning abilities across a wide range of medical subjects and topics.

Baselines We consider 4 baseline methods in our experiments, including Lexical Similarity (LS) (Fomicheva et al., 2020), Semantic Entropy (SE) (Kuhn et al., 2023), Predictive Entropy (PE) (Kadavath et al., 2022) and Length-normalized Entropy (LE) (Malinin & Gales, 2021). **Lexical Similarity** among a set of generated texts is quantified by computing the average ROUGE-L score, and a higher similarity score indicates that the model is more certain in its responses. **Semantic Entropy** addresses the difficulty of semantic equivalence in the uncertainty estimation of free-form LLMs by clustering generations with the same semantic meanings and calculating cluster-wise entropy. **Predictive Entropy** is estimated by averaging the sum of negative log probabilities of each token in the sampled answers for a given question. **Length-normalized Entropy** divides the sum of negative log probabilities