Prompts of Inconsistent Reasoning Attacks

1 Mutation

1.1 Paraphrase Attack

Input: {input_claim}
Task: Consider the above claim and paraphrase it while preserving its meaning.

1.2 Union Attack

Input: {input_claim} and {DB_claim}
Task: Make a union of these two claims and create a new claim while preserving semantics.

1.3 Negation Attack

Adopt or use the following examples:

Original: "Adult tissue-resident macrophages possess a self-renewing capacity."
Negation: "Adult tissue-resident macrophages do not lack a self-renewing capacity."

Original: "All six cathelicidins are located on chromosome two."
Negation: "None of the six cathelicidins are not located on chromosome two."

Original: "Bronchial responsiveness is higher in the summer than in the winter."
Negation: "Bronchial responsiveness is not lower in the summer than in the winter."

Original: "Carriers of the alcohol aldehyde dehydrogenase deficiency mutation drink more than non-carriers."
Negation: "Carriers of the alcohol aldehyde dehydrogenase deficiency mutation do not drink less than non-carriers."

Original: "Choice of specialty is influenced by debt accrued from medical training."
Negation: "Choice of specialty is not unaffected by debt accrued from medical training."

Original: "Excess nitric oxide is responsible for vasospasm."
Negation: "Excess nitric oxide is not without responsibility for vasospasm."

Original: "Genomic instability in leukemia cells results from a decrease in reactive oxygen species from oncogene activation."
Negation: "Genomic instability in leukemia cells does not result from a lack of decrease in reactive oxygen species from oncogene activation."

Original: "Incidence rates of cervical cancer have increased over time."
Negation: "Incidence rates of cervical cancer have not failed to increase over time."

Input: {input_claim}
Task: Convert the claim using the negation technique while preserving its semantics.

1.4 General-to-Specific Attack

Given a claim, the general version of that claim converts a specific term to its corresponding general term. Conversely, the specific version of a claim converts a general term to a corresponding specific term.

Following the examples given below, convert a given claim into its specific version:

Original: "Cancer can be treated with surgery."
General-to-Specific: "Breast cancer can be treated with surgery."

Original: "Blood cells play a role in immunity."
General-to-Specific: "White blood cells play a role in immunity."

Original: "Bariatric surgery is used for severe obesity."
General-to-Specific: "Roux-en-Y gastric bypass is used for severe obesity."

Original: "White blood cells increase during bacterial infections."
General-to-Specific: "Neutrophils increase during bacterial infections."

Original: "Bleeding events can occur after surgery."
General-to-Specific: "Nosebleeds can occur after surgery."

Original: "Birds migrate during winter." General-to-Specific: "Sparrows migrate during winter."

Original: "People can suffer from chronic illnesses."
General-to-Specific: "Women can suffer from chronic illnesses."

Original: "Breast cancer has a poor prognosis."
General-to-Specific: "Triple-negative breast cancer has a poor prognosis."

Original: "Primary bile acids help in digesting fats."
General-to-Specific: "Chenodeoxycholic acid helps in digesting fats."

Input: {input_claim}
Task: Convert the given claim into its specific version.

1.5 Vague Attack

Given a claim, the vague version of that claim converts a specific term to its corresponding vague term. For example: "suppresses" to "affects", "reduce" to "affect", "positively associated" to "associated" etc.

Following the examples given below, convert a given claim into its vague version:

Original: "Activation of PPM1D suppresses p53 function."
Vague: "Activation of PPM1D affects p53 function."

Original: "White Blood cells are an important part of your immune system, which helps to detect and deal with infections."
Vague: "White blood cells are an important part of your immune system, which interacts with infections."

Original: "Antibiotic induced alterations in the gut microbiome reduce resistance against Clostridium difficile."
Vague: "Antibiotic induced alterations in the gut microbiome affect resistance against Clostridium difficile."

Original: "Antiretroviral therapy reduces rates of tuberculosis across a broad range of CD4 strata."
Vague: "Antiretroviral therapy affects rates of tuberculosis across a broad range of CD4 strata."

Original: "Asymptomatic visual impairment screening in elderly populations does not lead to improved vision."
Vague: "Asymptomatic visual impairment screening in elderly populations does not affect vision."

Original: "Auditory entrainment is strengthened when people see congruent visual and auditory information."
Vague: "Auditory entrainment is affected when people see congruent visual and auditory information."

Original: "Bariatric surgery has a positive impact on mental health."
Vague: "Bariatric surgery has an effect on mental health."

Original: "Birth-weight is positively associated with breast cancer."
Vague: "Birth-weight is associated with breast cancer."

Original: "Primary bile acids help in digesting fats."
Vague: "Primary bile acids impact in digesting fats."

Consider the given claim: {input_claim}.
Task: Convert it to vague.

1.6 Generalization Attack

Given a claim, the general version of that claim converts a specific term to its corresponding general term. Conversely, the specific version of a claim converts a general term to a corresponding specific term.

Following the examples given below, convert a given claim into its general version:

Original: "Breast cancer can be treated with surgery."
Generalization: "Cancer can be treated with surgery."

Original: "White blood cells play a role in immunity."
Generalization: "Blood cells play a role in immunity."

Original: "Roux-en-Y gastric bypass is used for severe obesity."
Generalization: "Bariatric surgery is used for severe obesity."

Original: "Neutrophils increase during bacterial infections."
Generalization: "White blood cells increase during bacterial infections."

Original: "Nosebleeds can occur after surgery."
Generalization: "Bleeding events can occur after surgery."

Original: "Sparrows migrate during winter."
Generalization: "Birds migrate during winter."

Original: "Women can suffer from chronic illnesses."
Generalization: "People can suffer from chronic illnesses."

Original: "Triple-negative breast cancer has a poor prognosis."
Generalization: "Breast cancer has a poor prognosis."

Original: "Chenodeoxycholic acid helps in digesting fats."
Generalization: "Primary bile acids help in digesting fats."

Consider the given claim: {input_claim}.
Task: Convert it to general.

2 Crossover

Input: {input_claim1} and {input_claim2}
Task: Combine these two claims and create a new claim while preserving their semantics.

3 Self-Reflection

Input: {input_claim}
Output: {output_claim}
Context: {context}

Task: If the output claim is consistent with the context and was correctly generated from the input claim, return **nothing** (empty response).
Otherwise, provide a short and concise **reflection/feedback** indicating why the output claim is incorrect or inconsistent with the context.

4 Retrieval-Augmented Generation (RAG)

Scientific evidence is below:
{context_str}

Given the context information and no prior knowledge, is the following claim **support** or **refute** based on the given evidence?
Claim: {input_claim}

5 Attack-Reflection-RAG

5.1 Attack Reflection Creation

Scientific evidence is below:
{context_str}

Some possible text attack scenarios:
{attack_descriptions}

Based on the given scientific evidence, attack description and without relying on prior knowledge, determine whether the claim: {input_claim} is under attack and classification {label} is correct or incorrect.

If **classification is correct**, return **nothing** (empty response).
If **classification is under attack or incorrect**, provide a short and concise **reflection/feedback** to aware the LLM.

5.2 Classify with Attack Reflection

Scientific evidence is below:
{context_str}

Your generated reflection:
{reflection}

Given the context information, reflection and no prior knowledge, is the following claim **support** or **refute** based on the given evidence?
Claim: {input_claim}

Reproducibility

All code for our approach, experiments, and parameter settings is available anonymously at: https://github.com/orca-great-white/inconsistency-reasoning.