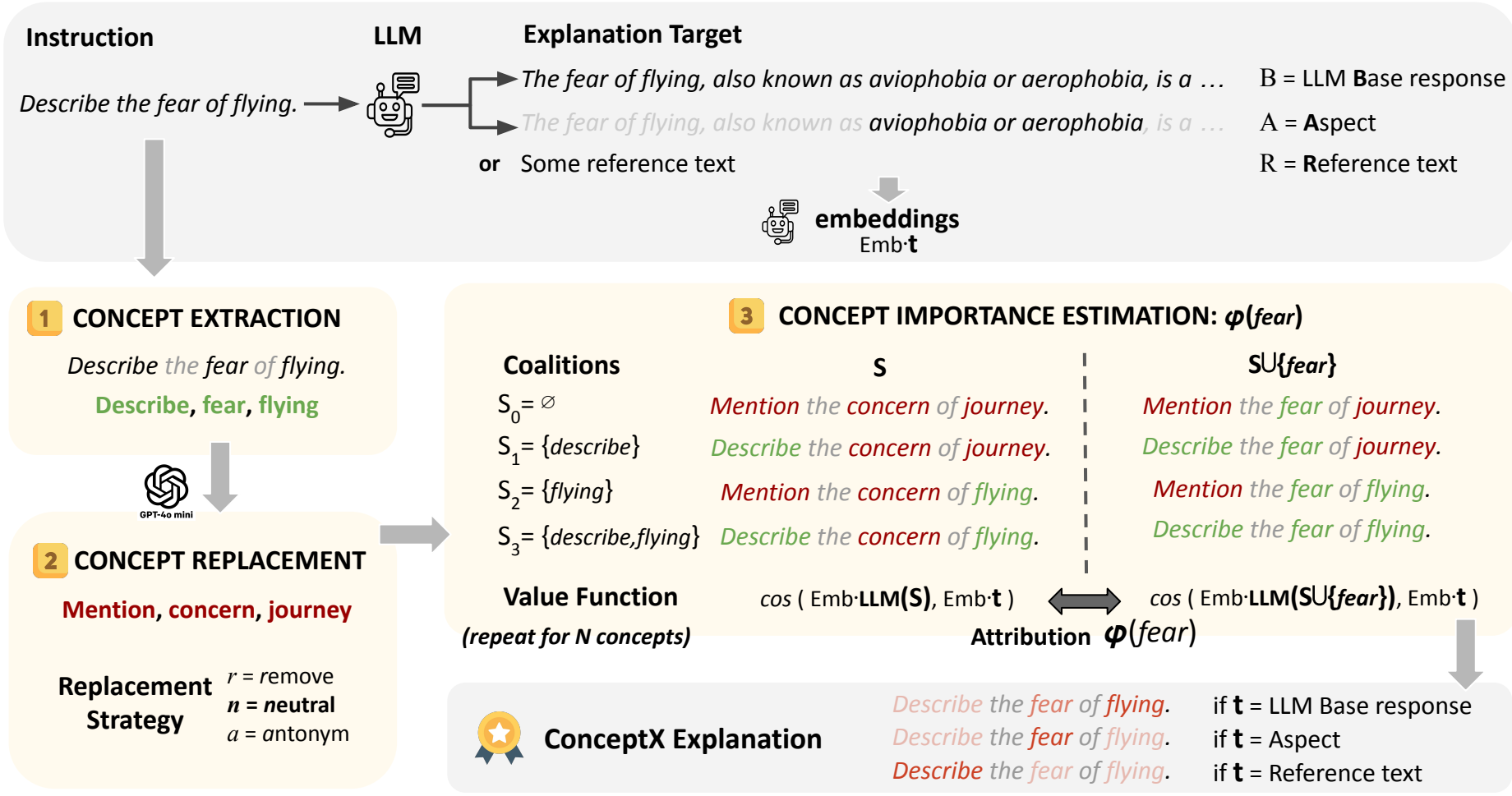
Instruction
LLM
Explanation Target
Describe the fear of flying.
The fear of flying, also known as aviophobia or aerophobia, is a …
The fear of flying, also known as aviophobia or aerophobia, is a …
or Some reference text
B = LLM Base response
A = Aspect
R = Reference text
embeddings
Emb·t
1 CONCEPT EXTRACTION
Describe the fear of flying.
Describe, fear, flying
GPT-4o mini
2 CONCEPT REPLACEMENT
Mention, concern, journey
Replacement Strategy
r = remove
n = neutral
a = antonym
3 CONCEPT IMPORTANCE ESTIMATION: φ(fear)
Coalitions
S
S∪{fear}
S0 = ∅
S1 = {describe}
S2 = {flying}
S3 = {describe,flying}
Mention the concern of journey.
Describe the concern of journey.
Mention the concern of flying.
Describe the concern of flying.
Mention the fear of journey.
Describe the fear of journey.
Mention the fear of flying.
Describe the fear of flying.
Value Function
(repeat for N concepts)
cos (Emb·LLM(S), Emb·t)
cos (Emb·LLM(S∪{fear}), Emb·t)
Attribution φ(fear)
ConceptX Explanation
Describe the fear of flying.
Describe the fear of flying.
Describe the fear of flying.
if t = LLM Base response
if t = Aspect
if t = Reference text