Agentic SelfCheck Framework

Semantic Agent

Word Embedding
Count N-grams
Tokenize
Cosine Similarity
Hallucination Score
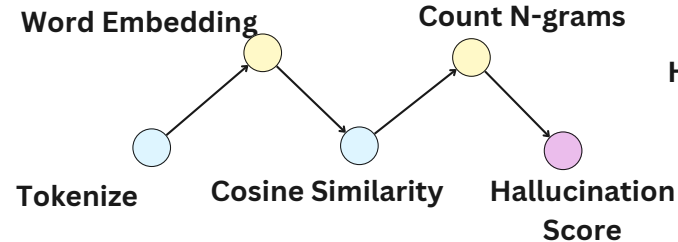
Premise + Hypothesis → ZS / CoT → Pretrained Model → Hallucination Score
Logits for NLI classes
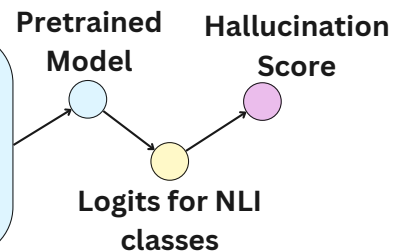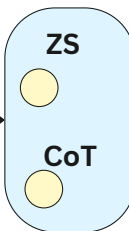
[The dog is happy,...
The canine is joyful]

Hallucination score : 0.10 ✓
Hallucination score : 0.63 ✗

Specialized Detection Agent

Phi-3
Mistral 7B
Llama 3.1 8B
Gemma 7B

→ Finetuning / MNLI → NLI Model → Hallucination Score

Premise + Hypothesis
Classofication to NLI classes
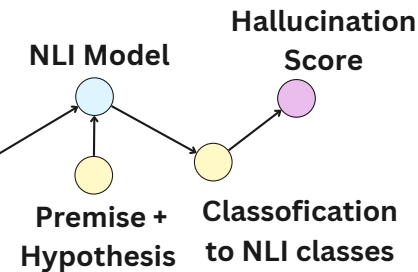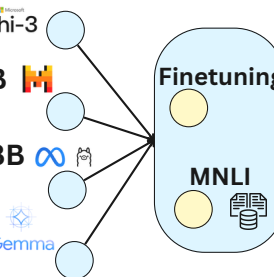
[Premise: The Earth is round.
Sentence: The Earth is flat.]

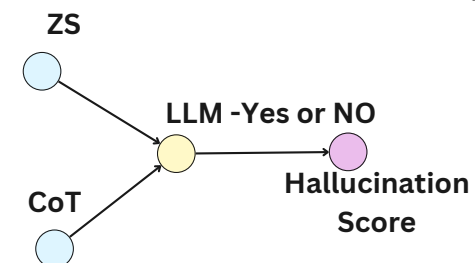Hallucination score : 0.89 ✓
Hallucination score : 0.49 ✗

[Premise: The cat is sitting on the mat.
Hypothesis :A cat is resting on the mat.]

Hallucination score : 0.05 ✓
Hallucination score : 0.71 ✗

Contextual Consistency Agent

ZS
CoT
→ LLM -Yes or NO → Hallucination Score

[Context: The sky is clear and sunny today.
Sentence: It is raining heavily outside.]

Hallucination score : 1.00 ✓
Hallucination score : 0.41 ✗