Limitations and Future Enhancements

Limitations:

- The credibility scoring relies on heuristic and metadata-based signals but does not deeply understand the semantic content of the claims that is crucial.
- It cannot detect contextually misleading claims if they are phrased confidently.
- The claim text is parsed with Regex which is pattern dependent.
- The system does not cross-verify claims with external, authoritative knowledge sources, does not cross-verify claims with external, authoritative knowledge sources.
- As a result, scores may inflate even if the claim is factually false but appears in a high-reputation source.
- Contradiction logic is simple. It misses nuanced contradictions such as causality reversals, misattributed data, or implicit refutations.
- Domain reputation and source weighting are manually defined and static, they don't evolve over time or based on real world statistical metrics.
- New or emerging sources will default to "unknown," limiting scoring accuracy.
- Incremental updates are score-based only, they don't adapt or learn from historical credibility outcomes or evolving evidence as no feedback loop or reinforcement mechanism is developed.
- The system outputs a score and action but does not provide a detailed explanation of why a claim was flagged or validated.

Future Enhancements:

- Integrate transformer-based models for deep semantic credibility scoring.
- Utilise spaCy, NER or LLM's for context extraction and more accurate claim text parsing.
- Flag misinformation, logical fallacies, and unsupported correlations beyond keywords and metadata via LLM's.
- Add automated fact-checking pipelines using APIs.
- Cross-reference numerical claims and named entities in real-time.
- Build an active learning self-updating source reputation pipeline that learns from historical credibility outcomes, external ratings, and user feedback.
- Support continuous ingestion of new sources without manual updates.
- Use natural language inference (NLI) models to detect semantic contradictions.
- Provide a human-readable explanation for each score.
- Incorporate SHAP algorithm to explain predictions.
- Add a traceable audit log of scoring decisions.
- Allow domain experts or moderators to provide feedback on claim credibility.
- Use feedback to fine-tune scoring models and reputation weights over time.