Claims:

1. A method for generating a confidence score specific to an output from an artificial intelligence system comprising:

processing an input query by a language model to obtain an initial output;

computing a plurality of confidence estimates for the initial output, wherein each confidence estimate of the plurality of confidence estimates is computed based on a distinct computational technique from a plurality of distinct computational techniques;

aggregating the confidence estimates based on a game-theoretical approach to generate an aggregated confidence score for the initial output; and

applying the aggregated confidence score to adjust parameters of an artificial intelligence system,

wherein the game-theoretic approach is derived from a Nash Embedding Theorem, which optimizes the aggregation of the confidence estimates by evaluating an interdependency among the plurality of confidence estimates to reach an equilibrium among computational strategy spaces of the plurality of distinct computational techniques.

2. The method of claim 1, wherein the adjusted parameters of the artificial intelligence system are based on model performance metrics corresponding to a level of confidence determined by the aggregated confidence score.

3. The method of any one of the preceding claims, wherein the distinct computational technique is selected from a group comprising: verbalized confidence assessments, token probability analysis, prompt entropy measurements, semantic output clustering, self-consistency diagnostics; and hidden state divergence metrics.

4. The method of any one of the preceding claims, further comprising using a structured data format to encode the confidence estimates and the aggregated confidence score, wherein the format is compatible with a standardized metadata representation.

5. The method of any one of the preceding claims, further comprising refining the game-theoretical approach based on feedback derived from system performance to enhance future confidence estimations.

6. The method of any one of the preceding claims, wherein the distinct computational technique includes calculating token probabilities by extracting SoftMax probabilities for each

token from an output distribution of the language model and calculating an overall sequence probability as a product of the token probabilities.

7. The method of any one of the preceding claims, wherein the distinct computational technique includes determining a prompt entropy by prompting the AI system with multiple label options, calculating an entropy of a resulting probability distribution corresponding to the label options, and mapping entropy values to the confidence estimates based on a predetermined function.

8. The method of any one of the preceding claims, wherein the distinct computational technique includes analyzing a verbalized confidence by pattern matching to identify statements of confidence in an output of an AI system and assigning numerical scores to the identified statements based on a predefined mapping.

9. The method of any one of the preceding claims, wherein the distinct computational technique includes clustering semantic outputs by generating a plurality of output sequences, computing semantic embeddings for the output sequences, and using metrics to quantify a clustering quality.