1. Problem Definition (6 points)

AI Problem:
"Detecting Fake News on Social Media Platforms"

Objectives:

1. To classify news articles/posts as either *real* or *fake* using natural language processing.
2. To alert platform moderators of highly probable fake news for further review.
3. To reduce the spread of misinformation by at least **30%** through early detection and intervention.

Stakeholders:

- ✓ **Social Media Companies** (e.g., Facebook, Twitter) who are concerned with maintaining credibility and user safety.
- ✓ **Government Agencies / Fact-checkers** who are responsible for monitoring public information and national security.

Key Performance Indicator (KPI):

F1 Score of the fake news classifier that balances false positives and false negatives in detection.

2. Data Collection & Preprocessing (8 points)

Data Sources:

1. **Fake News Dataset from Kaggle:** This contains labeled news headlines and articles tagged as real or fake.
2. **News APIs** like **NewsAPI.org:** Used to fetch live news from trusted sources for comparison and evaluation.

Potential Bias:

Source Bias: If most training data comes from Western or English-speaking sources, the model may struggle with non-Western or multilingual misinformation, leading to geographic/cultural skew.

Preprocessing Steps:

1. **Text Cleaning:** Removing HTML tags, URLs, special characters.
2. **Tokenization & Stopword Removal:** Breaking text into words and remove common words like "the", "and".
3. **Vectorization:** Converting text into numerical format using TF-IDF or Word Embeddings for model input

3. Model Development (8 points)

Chosen Model is:

Bidirectional LSTM (BiLSTM) : This is a type of recurrent neural network that understands word context from both directions, ideal for nuanced text like news articles.

Justification:
It captures sequential context in text better than traditional models like Naive Bayes or Logistic Regression. BiLSTM performs well on NLP tasks such as classification and sentiment analysis.

Data Splitting Strategy:

70% Training, **15% Validation**, **15% Test**

Stratified sampling is used to maintain balanced real/fake class distribution in all sets.

Hyperparameters to Tune:

i. **Learning Rate:** Controls how fast the model adjusts weights during training.
ii. **Batch Size:** Affects memory usage and learning stability during training.

4. Evaluation & Deployment (8 points)

Evaluation Metrics:

i. **F1 Score:** Ensures balance between Precision (false positives) and Recall (false negatives) in detecting fake news.
ii. **AUC-ROC Curve:** Measures the model's ability to distinguish between fake and real news across all thresholds.

Concept Drift:

Definition: A situation where the statistical properties of incoming data change over time (e.g. new types of misinformation arising).

Monitoring Strategy:

- ✓ Regular re-evaluation of model on recent data.
- ✓ Implement **drift detection tools** like Alibi Detect, Amazon SageMaker Model Monitor.
- ✓ Periodically retrain the model on updated datasets.

Technical Challenge in Deployment:

Scalability: Serving real-time predictions for millions of social media posts per day demands scalable infrastructure e.g., Dockerized models with load balancing on Azure or AWS Lambda.