📋Report: Hybrid Grammar-Scoring Engine

---

1. Overall Approach

Goal: Predict a continuous grammar-proficiency score (0–5) for spoken responses.

Why hybrid?

- Pure deep-learning (e.g. BERT alone) can miss explicit error counts that humans notice.
- Pure feature-based ML can't fully capture context or subtle usage patterns.

Components:

1. **Hand-crafted NLP features**
    - These are interpretable signals—e.g. "this response has 12 grammar errors," or "average sentence length is 8 words."
    - They directly quantify aspects of grammar that human raters consider.
2. **Classical ML ensemble**
    - RandomForest, LightGBM, and Ridge each learn different patterns in the feature space.
    - Averaging them reduces overfitting and stabilizes predictions.
3. **DistilBERT regression**
    - Leverages large-scale pretraining on language to understand context, agreement, and nuance.
    - Fine-tuned to map cleaned transcript text → grammar score.
4. **Meta-stacking**
    - A simple Ridge model learns how to weight "feature_pred" vs. "bert_pred."
    - Ensures that when one branch is uncertain, the other can compensate.

---

2. Preprocessing Steps

Every stage ensures data quality and extracts maximum signal:

1. **Audio Cleanup**
    - **Resample to 16 kHz**: standardizes input for ASR.
    - **Normalize amplitude**: avoids variation in loudness skewing ASR confidence.
    - **Trim silence**: focuses transcription on speech, removes long pauses.
2. **ASR Transcription**
    - **Whisper-base** chosen for its balance of accuracy and speed.

    - Produces raw text, including disfluencies and filler words.
3. **Transcript Cleaning**
    - **Lowercasing**: reduces vocabulary size, avoids spurious mismatches.
    - **Remove fillers** (“um,” “uh,” “like”): these do not reflect grammar competence.
    - **Fix spacing/punctuation**: ensures downstream NLP tools (LanguageTool, spaCy) analyze correctly.
4. **NLP Feature Extraction**
    - **LanguageTool**: off-the-shelf rule-based grammar checker; counts errors like subject–verb disagreement.
    - **spaCy**:
        - Sentence segmentation → average sentence length (complex sentences often indicate higher skill).
        - POS tagging → diversity of parts-of-speech (varied vocabulary and structure).
    - **Normalization**: errors per word accounts for response length; interaction term (sentence_length × errors) captures whether long sentences incur more errors.
5. **GEC-Based Features**
    - Use a **T5 grammar-correction model** to propose corrected text.
    - **Levenshtein edit distance** between original vs. corrected text approximates how many corrections were needed.
    - **Edit rate** (edits ÷ words) normalizes for response length.

Outcome: A rich feature vector capturing both quantity (error counts) and quality (complexity, edit rates).

---

3. Pipeline Architecture

```
Raw Audio
   ↓ Audio Cleanup (resample, normalize, trim)
   ↓ Whisper ASR
Raw Transcripts
   ↓ Transcript Cleaning (remove fillers, punctuation)
   ↓─── Parallel Branch ─────────────────────────────┐
   │                                                  │
   │  Hand-crafted NLP & GEC features → Classical ML →│ feature_pred
   │                                                  │
   └─> DistilBERT regression on cleaned text → bert_pred
      (fine-tuned with warmup, weight-decay, fp16,
       best-model checkpointing on Pearson)
   ↓
Stack [feature_pred, bert_pred]
   ↓
Meta-regressor (Ridge) → Final Grammar Score (0-5)
```

---

4. Fine-Tuning Enhancements

To squeeze maximal correlation from DistilBERT:

- **Seeding and determinism**: ensures results repeat across runs.
- **Freezing lower layers**:
  - Lower layers encode general language patterns; freezing them prevents overfitting on small data.
  - Only higher layers adapt to grammar-scoring task.
- **Weight decay (0.01)** & **warmup (10%)**:
  - Weight decay regularizes model weights, reducing overfitting .
  - Warmup gradually ramps up learning rate, stabilizing early training.
- **Mixed precision (fp16)**: faster training and less memory usage without accuracy loss.
- **Checkpoint on best Pearson**: directly optimizes for the metric of interest, not just loss .

---

5. Evaluation Results

Metric	Value	Interpretation
RMSE	0.7190	On a 0–5 scale, average error ≈ 0.72 points.
Pearson r	0.5873	Moderate-strong linear correlation with human scores.

- **OOF residuals** are roughly Gaussian, indicating no major bias.
- **Pearson 0.587** shows the model captures relative ordering of responses well.

---

6. Future Work

1. **Audio-fluency features**
    - Pause durations, speech rate, jitter—these often correlate with proficiency.
2. **Multimodal fusion**
    - Combine raw audio embeddings (wav2vec2) with text features in a single network.
3. **Hyperparameter tuning**
    - Use Bayesian search for LightGBM and meta-regressor weights.
4. **Data augmentation**
    - Synthetic disfluencies or back-translation to enlarge training set.

Conclusion: This detailed pipeline—combining interpretable features, classical ensembles, and fine-tuned transformers—provides a robust, extensible solution for automated grammar scoring of spoken English.