# Helping an LLM improve its detection of TV Commercials in Sports Broadcasts

*Brad Brown (bradb416@stanford.edu)*

*Department of Computer Science, Stanford University*

Stanford
Computer Science

## Project Overview

Can the topic analysis capabilities of large language models be used for television commercial detection?

- SCOPE: **Sports broadcasting on television** in the USA. The high-level goal was to **detect commercials** within short segments of the broadcast.
- PROCESS: A promising role for LLMs in this field, but **required additional 2nd stage modeling** techniques. No existing research has tried LLMs.
- RESULTS: Initial results detecting commercials given **15 seconds** of TV sports broadcast transcripts (text) gives a balanced **accuracy score of 79%, which were boosted to 87% via logistic classification.**
- ISSUES: 1) F**alse positives** 2) Previous detection methods using visual analysis reach 96% accuracy.

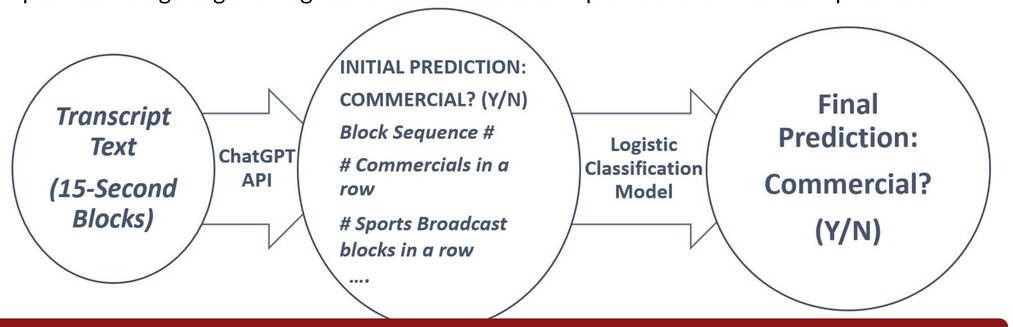
## Datasets & Metrics

- LLM prompts and the transcript text blocks are the key input data. The transcripts included **every spoken word** (not identifying the speaker). They included any spoken words from commercials as well. There were stored as 15-second elapsed time chunks in a custom SQL database schema
- My manual annotations captured one of 5 possible labels for each 15-second block:
  - Sports Broadcast
  - Sports Broadcast to Commercial
  - Commercial
  - Commercial to another Commercial
  - Commercial to Sports Broadcast
- Also kept a simple binary **Commercial Detected:**
  - **Yes/No Label**
- A total 2,891 15-second text blocks were captured in this project, representing **~722 minutes of video** across 2 professional basketball games and 2 professional football games, one involving double over-time.

## Methods & Experiments

The research work flow involved two high-level stages:

- Stage 1: The LLM prediction stage was to design LLM instruction prompts to classify chunks of text, feed those chunks to LLM API, and record results vs annotated labels.
- Stage 2: Final stage was to incorporate the LLM predictions, the 'streaks' of the presence (and sustained absence) of commercials and sports broadcasting blocks. Also, incorporated the number of blocks of text of each type seen so far. Together they formed the data input to post-LLM stage Logistic Regression SAGA model to improve on the initial LLM prediction.




## Discussions & Future Research

### Discussions:

- Logistic Classification Binary model measurably improved the accuracy - see "Classification report". Note the 4 and 8 point increases in accuracy and balanced accuracy to 91% and to 87%.
- The model with C = 1 regularization does not overfit ('Training Size vs Misclassification")
- False positives in commercial detection in both stages are a concern (~8%)
- The LLM Stage 1 and the LR Multinomial model (Stage 2) did NOT succeed in classifying transitions and combinations of events **within** the 15-second window. 37% Balanced Accuracy.
- No obvious game timeline pattern for accuracy of prediction - see "Commercials in a game"
- Insignificant differences between 2 experiment types: 15-second window/15-second context vs 15-second window/30-second context experiments.
- Best-performing LLM prompt instructed to predict categories, give its rationale for the answer, and make final second prediction within the same instruction, looking at its rationale.

### Future Research:

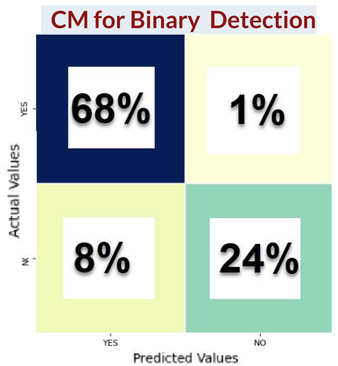
1) Systematically engineer and test additional prompts with full-scale testing vs a baseline
2) Capture the confidence or probability of LLM in its responses so I can better calibrate Stage 1
3) Try XGBoost and Poisson Point Process model as classifiers for the second stage
4) Develop an **ensemble approach with existing commercial detection techniques** such as black frame detection and volume variance
5) Dynamically change prompt to LLM as it is going through the event. Let it know what the second stage model thinks of its last prediction and also give it similar sequence data such as 'You predicted 4 commercial blocks in a row so far'. **Given the massive "memory" being added to LLMs, can give it the entire annotated training set from previous games. It is possible we don't need a second stage model.**

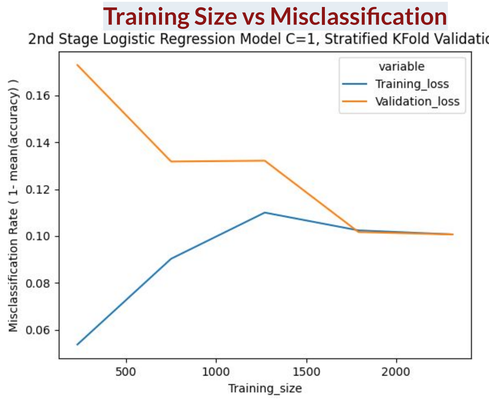
### References

[1] R. Lienhart, C. Kuhmünch and W. Effelsberg. On the Detection and Recognition of Television Commercials. Universität Mannheim Praktische Informatik IV L15.16 D-68131 Mannheim

## Results

| Classification Report | Stage 1-->Stage 2 = Gain |
|---|---|
| Accuracy | 0.87 --> 0.91 = +.04 |
| Balanced Accuracy | 0.79 --> 0.87 = +.08 |
| Precision (macro avg) | 0.91 --> 0.93 = +.02 |
| Recall (macro avg) | 0.79 --> 0.87 = +.08 |
| F1-Score (macro avg) | 0.82 --> 0.89 = +.07 |




***91% accuracy with simple Log Reg model. Future research could yield even better results.***