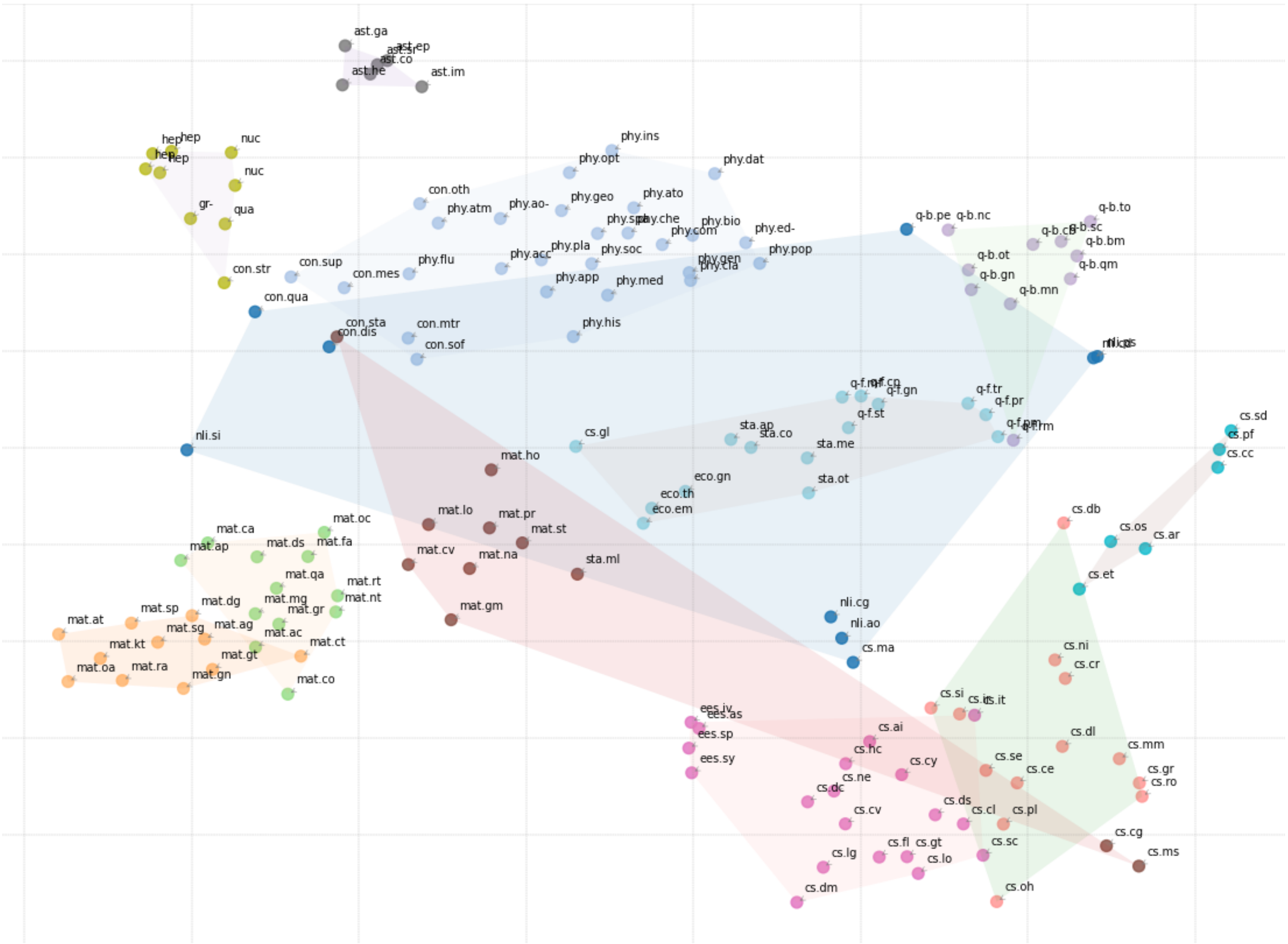

PROJECT 3
DEVELOPMENT OF A CASE STUDY WITH MODEL EVALUATION

Generative Language Models

António Cruz (140129), **Erik Daskalyuk** (120062), **Ivan Magalhães** (106586), **Ricardo Kayseller** (95813), **Ricardo Pereira** (120052)

Table of Contents

1. Problem Statement and Objectives

The project aimed to develop and assess an automatic classifier for arXiv papers, utilizing only the metadata available at the time of submission: the title and abstract. The concrete task was to predict the arXiv primary subject among 148 possible categories, covering all major scientific domains (computer science, mathematics, physics, statistics, quantitative finance, etc.).

The business motivation was two-fold:

- Reduce the manual effort and inconsistency involved in assigning subjects to new submissions.
- Provide a principled way to prioritize human review by surfacing high-confidence predictions and flagging low-confidence or ambiguous cases.

Although the broader vision is genuinely multi-label (papers often belong to several fields), the implementation in this case study focuses on predicting a single primary subject per paper, while still analyzing performance hierarchically at the coarser “domain” level and via top-k and confidence-based views. This reflects a realistic first production step: getting the main label right and understanding when we can trust the model.

Notes on computational constraints and limitations of the experimental setup

Although we had access to some high-end consumer GPUs and even a data center GPU with 96 GB of VRAM, the combination of a relatively large dataset and a long-context transformer model (Longformer, supporting a max_length=4096 token sequences) imposed substantial computational and runtime costs.

In practice, most of our training attempts were interrupted by CUDA out-of-memory errors, indicating that the effective capacity of the hardware was exceeded under certain configurations that would benefit from Longformer capabilities (e.g., larger max_length, batch sizes, or more complex training loops). This has constrained our ability to perform an extensive hyperparameter search (e.g., systematic exploration of learning rates, batch sizes, regularization), to run multiple random seeds for robustness, or to compare a wider range of architectures and training regimes.

Consequently, instead of Longformer, we opted for SciBERT, a variant of BERT (max_length=512) specifically designed for scientific text. This model change enabled a more stable training environment.

2. Data Understanding and Preparation

The work begins from the full arXiv “train.parquet” file, with around 2.55M papers. The exploration analysis revealed several important properties that directly shaped the modeling choices.

Document structure and length. Titles are typically short (around 8-12 words), while abstracts cluster around 100-200 words. Only a small fraction of abstracts (~1.4%) show signs of truncation. This supports the decision to rely primarily on title + abstract as the text signal: they are short enough for efficient transformer processing, but long enough to encode meaningful subject information.

Scientific markup and noise. A substantial proportion of abstracts contain LaTeX and HTML artefacts: inline math, environments, HTML tags, entities such as "&amp;", and Greek commands. These elements are essential for human readers but tend to fragment the tokenization and add noise for a pretrained encoder. This led to a light-weight cleaning strategy in the Data Preparation phase:

- Decode HTML entities.
- Strip raw HTML tags.
- Replace LaTeX math segments with a neutral placeholder [MATH].
- Normalize whitespace.

Importantly, no stemming, lemmatization or stop words removal is applied. The exploratory work and the literature review emphasize that transformer models like BERT and SciBERT expect natural text and are already robust to function words and morphology; aggressive normalization would risk a distribution shift relative to pretraining.

Label distribution and multi-label nature. The subject distribution is highly imbalanced at the corpus level: domains such as cs, math and cond-mat dominate, while many fine-grained labels are rare. In addition, the comparison between "primary_subject" and the multi-valued subjects field shows that only about half the papers are strictly "single label"; roughly 45% mention additional subjects beyond the primary one.

This confirms that the ground truth primary subject is only one facet of a richer multi-label reality, and that any single-label classifier is solving a deliberately simplified but still useful problem.

To make the modeling tractable while preserving diversity, the training dataset is restricted and balanced as follows:

- Only categories with at least 100 papers are kept.
- Per category, a maximum of 1,000 samples is drawn (stratified).
- This yields 148 categories, each with enough examples for meaningful training, and removes the extremely long tail of ultra-rare subjects.

The resulting balanced dataset is then stratified into train, validation and test sets (≈70/15/15), preserving label distribution across splits.

Text construction. For each paper, a single input string is built:

text = [TITLE] [SEP] [ABSTRACT]

In the best model, the configuration uses USE_TITLE = True and USE_CLEAN_TEXT = True, so the model sees both fields, lightly normalized, separated by an explicit marker.

3. Modeling Choices and Architecture

The core model is based on SciBERT ("scibert_scivocab_uncased"), a domain-adapted BERT variant pretrained on a large corpus of scientific articles. This choice is aligned with the task: SciBERT's vocabulary and representations are tuned to the style and terminology of scientific writing, unlike generic BERT models.

On top of SciBERT, the project implements an enhanced classification head:

- Obtain either the pooler output or the [CLS] token
- Apply LayerNorm, a fully connected projection, GELU activation and dropout.
- Feed into a final linear layer that outputs logits over 148 classes.

This head is simple enough to train reliably, but expressive enough to re-shape SciBERT's embeddings for the specific classification task.

Several optimization techniques are integrated:

- **Class weights** computed from the training distribution are applied to rebalance the loss, preventing frequent categories from dominating.
- **Focal loss with label smoothing.** The loss function combines focal scaling with smoothed targets. Focal loss (γ = 2) down-weights easy, already-correct examples and emphasizes hard cases. Label smoothing (α = 0.1) prevents the model from becoming over-confident, by distributing a small portion of probability mass away from the true class across the remaining 147 labels. Together, these mechanisms aim to improve generalization and calibration, especially on rare or ambiguous subjects.
- **Mix-up augmentation.** From epoch 3 onwards, mix-up is applied at the embedding level: pairs of examples are linearly interpolated, and the loss is computed against the corresponding convex combination of labels. This creates intermediate "virtual" samples between classes and encourages the classifier to behave more smoothly in representation space.
- **Differential learning rates.** The SCIBERT backbone is updated with a conservative learning rate (2e-5), while the classification head learns at 10 times that rate, allowing the new layers to adapt quickly without destabilizing the pretrained encoder.
- **Learning-rate scheduling and early stopping.** A linear warm-up schedule with 500 warm-up steps and early stopping with patience 6 prevents over-training. Although the configuration allows up to 40 epochs, training in practice stops after 12 epochs when validation accuracy ceases to improve.

The net effect is a model architecture that stays close to well-tested BERT fine-tuning practices, but with targeted improvements to handle label imbalance, over-confidence and the large number of classes.

4. Quantitative Results

4.1 Overall and Baseline Comparison

On the held-out test set, the best model achieves:

- Overall accuracy: **59.78%**.
- High-confidence accuracy (for predictions above the 0.7 threshold): **71.87%**.
- High-confidence coverage: **64.52%** of test predictions.

Relative to a previously established baseline of roughly 24% accuracy on the same 148-way task, this corresponds to an improvement of about **2.5×** in accuracy. This is a large gain, especially considering the difficulty of the problem and the modest per-class sample sizes after balancing.

Training curves show the expected pattern: training accuracy rises steadily beyond 95%, while validation accuracy plateaus around 60%. This indicates some overfitting, which is not surprising given the relatively small balanced dataset and the model capacity. However, early stopping, dropout, label smoothing and mix-up, keep the validation accuracy stable and prevent catastrophic overfitting.

4.2 Hierarchical Performance: Domains vs Fine Subjects

Because arXiv subjects have a natural hierarchy, the evaluation explicitly separates coarse domain performance from fine-grained category performance.

- At **domain** level (e.g., cs, math, astro-ph), the model reaches **82.91%** accuracy on the test set.
- At **fine subject** level (148 categories), the accuracy is the global **59.78%** quoted above.

The hierarchical breakdown is particularly informative:

- In **59.78%** of test cases, the model correctly predicts the fine subject.
- In an additional **23.12%** of cases, the fine subject is wrong, but the coarse domain is still correct.
- Only **17.09%** of test examples are assigned to an incorrect domain.

Taken together, this means that in around **82-83%** of cases, the model correctly identifies the broad research area of the paper, and in roughly **60%** of cases it also gets the exact arXiv subcategory right.

From a business perspective, this is valuable even when the fine label is occasionally off: for routing, triage, and search indexing, knowing that a paper belongs to “computer science machine learning” rather than “condensed matter physics” is often the most critical decision.

4.3 Per-Category Behavior and Production Readiness

The per-class classification report confirms that performance is highly uneven across the 148 categories, which is expected given the diversity of topics and varying degrees of linguistic distinctiveness.

- The best subjects, such as **High Energy Physics – Lattice (hep-lat)**, **Earth and Planetary Astrophysics (astro-ph.EP)**, **Accelerator Physics (physics.acc-ph)** and **Superconductivity (cond-mat.supr-con)**, achieve F1-scores in the **0.80-0.88** range with balanced precision and recall.
- A total of 30 categories surpass an F1-score of 0.70, many of them aligned with clear, well-defined scientific communities (e.g., cs.DL, cs.IR, math.OA, econ.EM).
- In contrast, a long tail of subjects which is often broad, interdisciplinary or under-represented, shows much lower F1-scores, sometimes below 0.3. These classes remain challenging.

To connect model performance to practical deployment, the notebook defines "production-ready" categories as those with F1 ≥ 0.75 on the test set and reports how many test samples fall within that subset. This yields a concise summary:

- A non-trivial number of categories satisfy a strict F1 threshold.
- These categories cover a substantial fraction of the test samples.
- They form a natural candidate set for early limited deployment, while other categories continue to be monitored or require further work.

The confusion-matrix heatmaps for the top and bottom categories make the error patterns visible: among the best classes, almost all mass sits on the diagonal with only minor leakage into conceptually neighboring labels; among the weakest, confusions are more diffuse and often involve semantically close subjects (e.g., different flavors of "applications" or "other statistics").

4.4 Confidence Calibration

Beyond raw accuracy, the project explicitly examines how predictive confidence correlates with correctness. The test predictions are bucketed into confidence intervals:

- <50%, 50-60%, 60-70%, 70-80%, 80-90% and 90-100%.

The empirical accuracy is computed in each bin. The results show a monotonic pattern:

- At confidences below 50%, accuracy is roughly **30%**.
- In the 70–80% bin, accuracy climbs to about **56%**.
- In the 90–100% bin, accuracy reaches about **80%**.

This confirms that the model's probabilities are reasonably calibrated: higher scores genuinely correspond to more reliable predictions. Combined with the high-confidence accuracy of ~72% at ~65% coverage, this yields a useful operating point:

- Accept predictions automatically above a given confidence threshold (e.g. 0.8-0.9).
- Send low-confidence cases to a human reviewer or a more expensive second-stage model.

This kind of confidence-aware workflow is precisely what the initial business framing envisioned.

5. Impact of Design Choices

Taken together, the experimental results illustrate the practical impact of the main design decisions:

- **Domain-specific encoder (SciBERT).** Using SciBERT rather than a generic BERT is consistent with the strong performance on physics, mathematics and computer science categories, even though the balanced dataset is relatively small. The encoder brings a great deal of prior knowledge about scientific language.
- **Light text cleaning.** The cleaning step removes HTML/LaTeX artefacts and normalizes whitespace while preserving the linguistic content. The model can focus on scientific phrases instead of token fragments like <, br, or half-parsed TeX commands. This is a low-risk, high-reward transformation.
- **Balanced sampling and class weights.** By enforcing minimum and maximum samples per category, the model avoids being dominated by the largest labels and learns meaningful decision boundaries for smaller but still frequent categories. The residual imbalance is compensated with class-weight loss.
- **Focal loss with label smoothing.** This combination explicitly targets two pain points of multi-class classification at scale: the dominance of easy negatives and the risk of over-confident Softmax outputs. The stable validation accuracy and the confidence/accuracy curves suggest that this design contributes to usable calibration and robustness.
- **Mix-up augmentation.** Mix-up starts after a few warm-up epochs, once the classifier has learned basic structure from real examples. While its direct quantitative contribution is not disentangled in ablation studies, the generalization gap between training and validation remains controlled despite the high training accuracy, which is compatible with mix-up helping to regularize the model.

Overall, the pipeline demonstrates that carefully adapting a single, well-chosen encoder model with robust training tricks can deliver competitive performance on a challenging 148-class scientific classification task, without resorting to more complex architecture.

6. Limitations

There are several important limitations that should be acknowledged explicitly:

1. **Single-label approximation.** The current model predicts only the primary subject, ignoring the multi-label nature of many papers. This is a deliberate simplification for the case study, but it means that the system cannot yet express that a paper simultaneously belongs to, for example, "cs.LG" and "stat.ML".
2. **Restricted training set.** To keep training feasible, the model is trained on a balanced subset with at most 1,000 examples per category. This sacrifices the tail of ultra-rare

classes and discards a large portion of the available data for the frequent ones. While this design improves per-class stability, it also sets an upper bound on what the model can learn about rare or subtle distinctions.

3. **Overfitting and long-tail performance.** Training accuracy approaching 96–97% indicates that the model fits the training set very well, while test accuracy stays at ~60%. Some of this gap is structural (the problem is genuinely hard), but some reflect overfitting. The long tail of low-F1 classes is where this overfitting and data scarcity are most visible.
4. **No explicit modeling of label correlations.** Although the EDA and confusion matrices clearly show that certain subjects co-occur in meaningful ways, the current Softmax head treats labels as mutually exclusive. True multi-label architectures (sigmoid outputs, multi-label losses) or label-graph regularization are not yet implemented.
5. **Limited ablation and robustness analysis.** Because of time and compute constraints, the final configuration (SciBERT + cleaning + focal loss + smoothing + mix-up) was not systematically compared against simpler variants in controlled ablations. The conclusions about individual techniques are therefore qualitative rather than strictly causal.

Despite these limitations, the system already delivers strong, usable performance for many categories and provides a clear path for iterative improvement.

7. Final Remarks

From the perspective of the course requirements, our case study covered the full CRISP-DM cycle on a real world, non-trivial problem, in several aspects:

- A clearl objective (automated subject classification for arXiv-like repositories).
- A rich, carefully analyzed dataset with documented artefacts and biases.
- A well-justified modeling approach centered on SciBERT and modern training techniques.
- A thorough evaluation using appropriate metrics (accuracy, F1 per class, hierarchical domain accuracy, confidence/accuracy curves).
- An honest interpretation of strengths, weaknesses and deployment readiness.

The final model does not solve arXiv classification once and for all, but it demonstrates that a single transformer-based system, trained on a balanced subset of the corpus with thoughtful regularization and evaluation, can already match or exceed a strong baseline by a wide margin and deliver practically useful predictions for a substantial portion of the label space.

In that sense, the project achieves its primary goal: to show, with quantitative evidence and a transparent methodology, how a modern language model can be applied to a real-world scientific classification task and how its performance can be rigorously evaluated and interpreted.