Justification for Single Model Deployment Strategy in Emotion Classification System

Breda University of Applied Sciences

Academy of Games and Media

Block D- Engineer – MLOps – Natural Language Processing

Louie Daans, Radna Puriel, Gin Li, Carmen-Nicola Ioniță, Zakariae el Moumni

Team – NLP3

Selection

The Single Model deployment strategy was determined to be the optimal approach for this natural language processing (NLP)-based emotion classification system based on the following factors:

1. Project Scope and Complexity

The application processes any type of video and uses the transcripts to classify emotions, a task performed by a single BERT-based model.

Unlike multi-model ensemble systems, this project does not require complex routing or version comparisons (Karau et al., 2022).

2. Resource Constraints

Academic projects typically operate under limited computational budgets (Microsoft Azure, 2023).

Running multiple models (e.g., Shadow or Blue-Green) would double cloud costs, violating fiscal constraints (in a real work environment).

3. Risk Assessment

The emotion classifier was rigorously validated during development (F1-score > 0.72 across emotion classes). This score exceeds the baseline of 0.70 commonly used for real-world affective computing applications (Akhtar et al., 2020).

4. Operational Simplicity

A single endpoint reduces monitoring overhead compared to Canary or A/B testing deployments.

Eliminates traffic-splitting complexity since all requests route to one model (Burns et al., 2019).

Comparative Analysis Against Alternative Strategies

To determine the most appropriate deployment method, we conducted a comparative evaluation of commonly used rollout strategies, summarized in Figure 1. Strategies such as Direct (Big Bang) deployment were ruled out as inapplicable, as our project did not involve replacing a pre-existing production model—this approach is typically reserved for version transitions. Shadow Mode, while useful in high-stakes environments to compare live and test outputs, was dismissed due to the high cost of maintaining duplicate GPU-backed instances (~€150/month), offering little benefit in a single-model academic scenario. Similarly, Blue-Green deployment, which involves maintaining two identical production environments to reduce downtime risk, was considered excessive given the non-critical nature of this demonstration system and the resulting doubling of infrastructure costs. Canary deployment, which allows for gradual traffic shifting between versions, was unnecessary since the emotion classification service processes media files in batch without live user interaction, making progressive rollout irrelevant. Finally, A/B Testing was excluded because it requires multiple active models and user-facing feedback loops—neither of which apply to our batch-oriented inference pipeline.

Given these constraints, Azure ML Managed Online Endpoints, paired with a Single Model deployment strategy, emerged as the most cost-effective and operationally appropriate choice. This approach enabled full control over the deployed environment, minimized complexity, and aligned with MLOps principles without breaching the defined resource budget.

Deployment Strategy	Why Not Suitable	Key Limitation
Direct (Big Bang)	Unnecessary for initial deployment; no predecessor model exists to replace	Designed for version upgrades only
Shadow Mode	Prohibitive cost (~$160/month for redundant GPU instances) without clear benefit	Requires parallel model execution
Blue-Green	Over-provisioning (2x environments) unjustified for non-critical academic demo	Doubles infrastructure costs
Canary	No user-facing components require gradual rollout; all episodes processed uniformly	Adds complexity without benefit
A/B Testing	No comparative models exist yet; feedback mechanism irrelevant for batch processing	Requires multiple production models

Figure 1: Comparison of alternative strategies

Implementation Considerations

Fail-safes are implemented using Azure Monitor and Application Insights to track inference latency and output drift. Upon threshold breach, the managed endpoint triggers automatic rollback per policy configuration (Microsoft Azure, 2023c).

Despite single-model deployment, Azure's auto-rollback was configured to revert if:

Inference latency exceeds 500ms (95th percentile)

Emotion distribution anomalies detected (KL divergence > 0.2)

Cost Efficiency:

- Leveraged Azure ML's pay-per-use billing, scaling to zero during idle periods (Microsoft Azure, 2023).
- Projected monthly cost: $82.94 (NC6s_v3 spot instances) vs. $300+ for multi-model strategies.

Conclusion

The **Single Model deployment strategy** was selected based on its strong alignment with the project's **limited scope**, **academic resource constraints**, and the need for **operational**

simplicity. While it inherently carries more risk compared to phased rollout strategies, this risk is effectively mitigated through:

- **Rigorous pre-deployment validation** (achieving precision and recall above 70%)
- **Automated rollback mechanisms** triggered by latency or anomaly thresholds
- **Execution during controlled, low-traffic processing windows**

This strategy strikes a practical balance between MLOps best practices and the project's academic limitations. It enables efficient, cost-effective deployment without sacrificing reliability or maintainability.

References

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., … & Zimmermann, T. (2019). Software engineering for machine learning: A case study. *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). Borg, Omega, and Kubernetes. *ACM Queue, 14*(1), 70–93. https://doi.org/10.1145/2898442.2898444

Karau, H., Warren, J., & Zaharia, M. (2022). *Machine learning systems design*. O'Reilly Media.

Microsoft Azure. (2023). *Managed online endpoints documentation*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/concept-endpoints

Microsoft Azure. (2023b). *Pay-as-you-go pricing*. https://azure.microsoft.com/en-us/pricing/details/machine-learning/

Microsoft Azure. (2025, March 5). *Perform safe rollout of new deployments for real-time inference*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-safely-rollout-online-endpoints?view=azureml-api-2&utm_source=chatgpt.com&tabs=azure-cli