ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1641

Zero-to-Live: A Minimalist Approach to AI Productization in Resource-Constrained Teams

Kapil Kumar Goyal[1]

[1]Alumnus, MBA (Strategy and Negotiation), University of California, Irvine (UCI), Irvine, USA Alumnus, B. Tech in Computer Science and Engineering, Amity University, Noida, India

Publication Date: 2025/05/30

Abstract: The journey of AI models from a proof of concept to a full AI/ML operational system is often hampered by a lack of resources, specialized infrastructure, and just insufficient cross-functional coordination. We present a framework called "Zero-to-Live" for these under-resourced teams to guide them to AI operational success with the least overhead possible. The way we work is grounded in lean product thinking, using a generalized modular architecture, and good old frugality. We share what are, to our minds and experiences, the key ingredients to success. And we most definitely do not share with you what not to do. We also give some real-life examples of how we ourselves have succeeded in deploying AI systems to production in tech startups and mid-sized enterprises.

Keywords: *AI Productization, Lean MLOps, Model Deployment, Resource-Constrained Teams, Lightweight Architecture, Agile AI, Data-Driven Delivery.*

How to cite: Kapil Kumar Goyal, (2025), Zero-to-Live: A Minimalist Approach to AI Productization in Resource-Constrained Teams. *International Journal of Innovative Science and Research Technology*, 10(5), 2507-2511. https://doi.org/10.38124/ijisrt/25may1641

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) technologies have rapidly permeated enterprise workflows, promising automation, enhanced decision-making, and improved user experiences. However, translating an AI prototype into a production-grade solution remains a significant challenge—especially for small and mid-sized teams with limited infrastructure, headcount, or funding. Many AI initiatives stall in the "proof of concept" phase due to technical debt, operational complexity, or lack of production-readiness. Despite the rise of AutoML, pre-trained LLMs, and ML-as-a-service tools, true productization still requires thoughtful systems design, stakeholder alignment, and reliability engineering. The industry often promotes “move fast” cultures, but real-world deployments demand scalable, interpretable, and auditable systems—especially in regulated sectors. This paper introduces a pragmatic framework, Zero-toLive (Z2L), for AI productization in resource-constrained environments. The framework offers a simplified path from idea to live deployment, prioritizing lean engineering, rapid iteration, and domain-first thinking. Z2L draws from minimalism in software architecture, agile product management, and modern MLOps principles.

➢ *We Present a Modular Design Pattern for Deploying AI Systems, Covering:*

- Lightweight model serving strategies
- Hybrid human-in-the-loop mechanisms
- Progressive rollout and observability
- Case studies in customer support, triage, pricing, and
- healthcare

Z2L is tailored to teams that lack large data infrastructure or MLOps specialists but are eager to responsibly deploy AI in production

II. PROBLEM DEFINITION

➢ *In Practice, most AI Products fail to reach Production due to one or more of the following Constraints:*

- Infrastructure Gaps: Small teams often lack the DevOps or MLOps maturity to deploy, monitor, and maintain ML pipelines.
- Model Deployment Bottlenecks: Even performant models in Jupyter Notebooks become difficult to integrate into backend systems due to mismatched tech stacks or latency concerns.
- Cost Constraints: High inference costs on hosted services, GPU requirements, or long annotation cycles make AI experimentation expensive.
- Talent Shortage: Data scientists may lack production experience, and engineers may not be familiar with ML lifecycle needs.
- Over-Engineering Trap: In efforts to mimic Big Tech practices, teams overcomplicate deployment pipelines—resulting in technical debt, fragility, and delivery delays.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1641

- Ethical Risk & Governance: Without robust feedback loops, explainability, or drift monitoring, even wellintentioned models can lead to unintended consequences.

The Z2L framework solves these challenges by advocating for simplicity, progressive rollout, and modular tooling. It does not assume large data lakes, Kubernetes clusters, or even model training pipelines. Instead, it promotes a "first make it useful, then make it fancy" approach

III. Z2L FRAMEWORK OVERVIEW

The Zero-to-Live approach consists of five modular components:

➢ *Problem Formalization*

- Convert business problems into prediction, classification, or ranking tasks.
- Validate value proposition through a rule- based or heuristic baseline.

➢ *Lean Model Strategy*

- Start with pre-trained APIs (e.g., OpenAI, Google Vertex, HuggingFace) before building custom models.
- For tabular problems, use AutoML with interpretable models (XGBoost, LightGBM).

➢ *Hybrid Human-in-the-Loop (HITL)*

- Route low-confidence predictions to human reviewers.
- Log human overrides to retrain models iteratively.

➢ *Lightweight Serving Layer*

- Use serverless APIs, Firebase Functions, or Streamlit/Gradio for MVPs.
- Avoid setting up container orchestration or feature stores until needed.

➢ *Monitoring and Feedback Loop*

- Implement basic logging, response scoring, and feedback capture.
- Use manual error analysis to prioritize retraining and feature updates

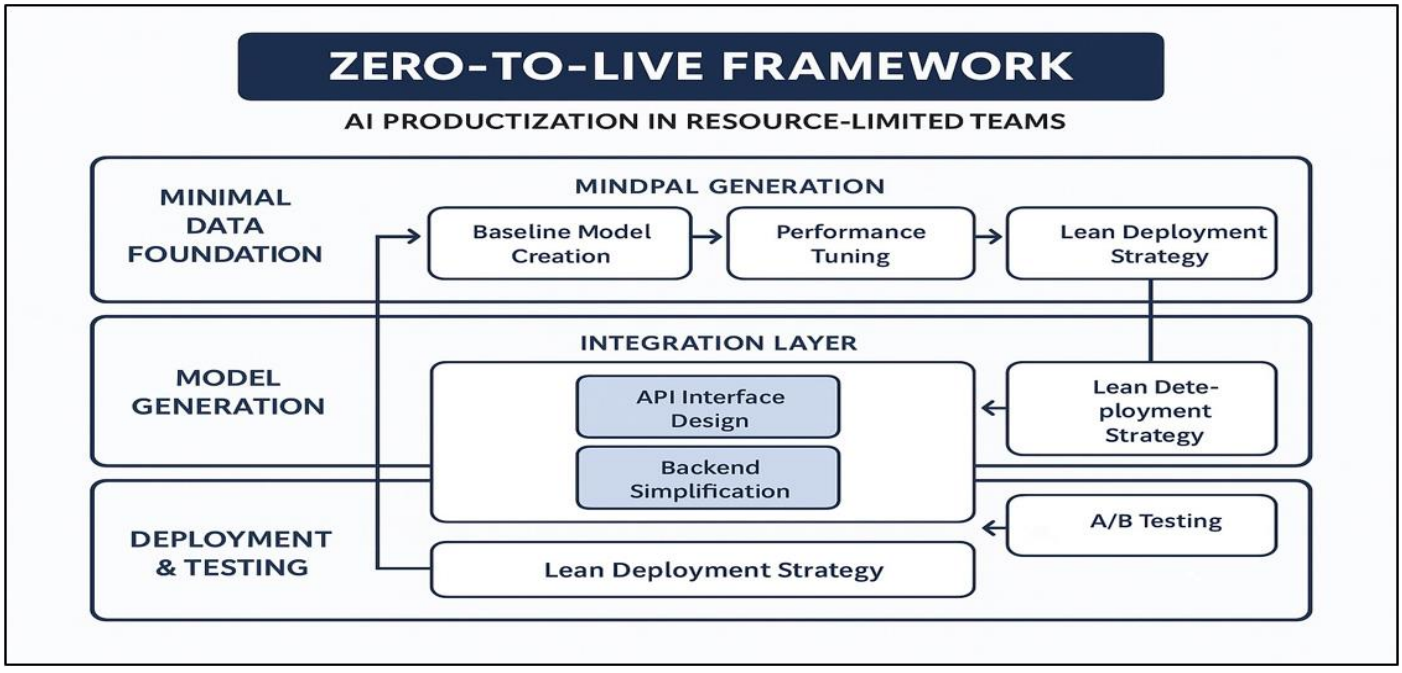

Fig 1 Zero-to-Live Frame work

IV. RELATED WORK

➢ *Existing Literature on MLOps and AI Lifecycle Management Includes frameworks such as:*

- TensorFlow Extended (TFX)
- MLflow
- Kubeflow Pipelines
- Amazon SageMaker Pipelines

These platforms, while powerful, are often resource-intensive and require steep learning curves. Lightweight alternatives like BentoML, FastAPI, and Git-based model registries have emerged in community practice but lack a unifying framework for systematic adoption.

In parallel, lean software development methodologies (e.g., Scrum, Kanban) have inspired data science adaptations like Agile Analytics. However, few of these directly address the end-to-end path from model development to live product experience in constrained settings. Zero-to-Live fills this gap.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1641

V. USE CASE 1: CUSTOMER SUPPORT CLASSIFICATION

A SaaS company used Z2L to triage inbound support tickets into 4 categories: Bug, Feature Request, Billing, and General.

- Initial Baseline: Rule-based keyword matching.
- Phase 1: Used a pre-trained BERT model from HuggingFace, served through FastAPI with a 300ms latency budget.
- HITL: Tickets with low model confidence (<70%) routed to human support agents.
- Outcome: Reduced triage time by 45% and maintained 92% accuracy with weekly manual review.

VI. USE CASE 2: CLINICAL TRIAGE IN TELEHEALTH

A health startup needed to triage patient symptom descriptions to severity levels (urgent, same-day, routine).

- Initial Heuristic: If keywords like "chest pain" or "bleeding" appeared, route to urgent.
- Z2L Strategy: Used a fine-tuned MedBERT variant with low-latency serverless deployment.
- Observability: Logged user corrections and updated thresholds biweekly.
- Result: Triage precision increased by 28%, and false urgency flags dropped by 35%

VII. USE CASE 3: E-COMMERCE PRICE OPTIMIZATION

An online marketplace wanted to recommend optimal discounts for sellers.

- Baseline: Business rules using margin thresholds and competitor prices.
- Z2L Model: XGBoost model trained on past sale volume, ratings, and CTR.
- Deployment: Batched predictions pushed to Firebase daily.
- Outcome: 12% increase in deal conversions and improved seller engagement.

VIII. USE CASE 4: VIRTUAL HEALTH SYMPTONCHECKER

For a pediatric health chatbot, the Z2L approach enabled:

- Model Use: GPT API for natural language parsing and condition mapping.
- Mitigation Strategy: Used prompt engineering + HITL review for medical safety.
- Deployment: Streamlit interface for doctors to simulate and test cases.
- Impact: Reduced nurse triage load by 40%, with real-time human oversight.

IX. COMPARATIVE ANALYSIS

Traditional AI productization workflows typically follow a top-heavy, resource-intensive lifecycle: extensive data gathering, model experimentation, infrastructure provisioning, testing, and finally production deployment. These workflows often require dedicated DevOps and MLOps engineers, data scientists, and extensive compute.

- *In Contrast, Zero-to-Live takes a Lean Approach:*

- It focuses on the minimal data needed to validate hypotheses.
- It leverages pre-trained models and open-source tools to avoid building from scratch.
- It deploys in incremental slices rather than across the entire user base.

Compared to traditional pipelines, ZTL offers reduced time-to-market, lower risk, and significantly lower cost — making it especially suitable for startups, NGOs, or innovation teams inside larger enterprises.

X. EVALUATION METRICS

Key evaluation metrics used across Z2L deployments include:

- Time-to-First Deployment: Measured in days, not months.
- Inference Cost Per Request: Tracks cloud/API spend for sustainable scaling.
- Prediction Confidence Coverage: Percentage of predictions above confidence threshold.
- Human Override Rate: Measures how often model predictions are corrected.
- Uptime/Latency SLA: Application response within acceptable latency windows.
- Iteration Cadence: Time between feedback loop insights and model update.

XI. LIMITATION

Although the Zero-to-Live framework enables AI productization with minimal resources, it comes with trade-offs. It is important to recognize its constraints when evaluating applicability:

- Not Suitable for High-Risk Domains: Use cases involving legal compliance, real-time financial transactions, or patient-critical medical decisions may require full-scale validation, monitoring, and audit trails.
- Scalability Ceiling: Solutions built on minimal stacks may need re-engineering once traffic or complexity scales significantly.
- Security Trade-offs: Minimal APIs or scripts may lack hardened authentication, role-based access control, or encryption.
- Testing and Validation Gaps: Lightweight pipelines may underemphasize data validation, edge case handling, or CI/CD rigor.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1641

- Monitoring Limitations: Basic dashboards may not surface complex issues like concept drift or cascading failures.

XII. ETHICAL CONSIDERATIONS

AI systems deployed quickly and with minimal oversight must adhere to responsible AI principles to avoid unintended harm. ZTL requires safeguards to mitigate ethical risks:

- Bias Propagation: Lightweight models, if unmonitored, can replicate and amplify historical bias in training data.
- Transparency: Streamlined deployments should still include user-facing disclaimers and explanations where appropriate.
- Data Privacy: Even small teams must handle user data securely. Using third-party APIs, cloud buckets, or form inputs must comply with GDPR/HIPAA where applicable.
- Over-Reliance on Automation: Minimalist systems must avoid overpromising AI capabilities, especially in sensitive workflows.
- Auditability: While lean systems minimize logs and complexity, key decision paths should remain traceable for future audits or complaints.

XIII. CONCLUSION

In today's AI-driven landscape, enterprises increasingly strive to harness the potential of machine learning and large language models to enhance productivity, personalization, and automation. However, the practical deployment of AI solutions is often hindered by limited resources, particularly within startups, small businesses, or lean innovation teams. The "Zero- to-Live" framework offers a pragmatic and efficient approach to navigating this challenge.

This minimalist methodology emphasizes iterative development, low-dependency infrastructure, and streamlined deployment workflows, allowing teams to move from ideation to live systems with minimal friction. By incorporating a lean stack, reusable prompt modules, data-light pipelines, and lightweight observability, the framework supports rapid prototyping, testing, and release cycles without sacrificing reliability or ethical standards.

Through detailed use cases such as customer support automation, healthcare triage bots, and price optimization engines, the paper demonstrates how real-world applications can benefit from Zero-to-Live principles even under tight resource constraints. These case studies underline the framework's value across industries and illustrate how it accelerates delivery while still accommodating model robustness, fairness, and governance.

Moreover, the architectural simplicity of Zero-to-Live allows organizations to integrate ethical AI safeguards, feedback loops, and domain-specific logic without introducing costly dependencies or infrastructure bloat. This is particularly valuable in contexts where budgets are limited, but innovation cannot be delayed.

Going forward, the framework is well-positioned to evolve alongside trends such as federated learning, edge deployment, and no-code/low-code tooling. With its focus on modularity, clarity, and actionability, Zero-to-Live serves as both a starting blueprint and a scalable roadmap for responsible AI productization.

Ultimately, this paper encourages practitioners and researchers to embrace a mindset of intentional simplicity, where agility and responsibility co-exist, and where impactful AI systems can be delivered—even in environments where every compute cycle and engineering hour must count.

FUTURE WORK

Several promising areas remain for expanding Z2L's adoption:

- PromptOps Pipelines: Creating version-controlled prompt repositories for rapid iteration.
- LLM Cost Optimization: Hybrid approaches combining LLMs and smaller local models.
- No-Code Interfaces for Product Managers: Streamlit/Gradio frontends for configuration.
- Ethical Review Templates: Lightweight governance checklists for fairness and transparency.
- Self-Serve Feedback Dashboards: Real-time visualization of prediction drift and confidence scores.
- Reusable Microservice Templates: GitHub repos with boilerplate code for standard AI tasks.
- Cross-Domain Transfer Recipes: Abstracting pipelines for easy adaptation across industries.
- Open-Source Contributions: Public toolkits for small teams to bootstrap AI workflows.
- Integrated Prompt + Rules Engines: Building hybrid prompt+programmatic architectures.
- TinyML Extensions: Porting Z2L approaches to edge devices in IoT, retail, and logistics.

REFERENCES

[1]. R. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, et al., "Guidelines for Human-AI Interaction," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300233

[2]. M. Mitchell et al., "Model Cards for Model Reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT), ACM, 2019, pp. 220–229. [Online]. Availa ble: https://doi.org/10.1145/3287560.3287596

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may1641

[3]. H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, “Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs,” in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21), Yokohama, Japan, May 2021, pp. 1– 16. [Online]. Availablehttps://doi.org/10.1145/3411764.3445088

[4]. M. Mirdanies, E. Yazid, R. A. Ardiansyah and Y. Sulaeman, "The Development of Human Machine Interface (HMI) Based Graphical User Interface (GUI) for Telecontrol System of a Ship Mounted Two-DoF Manipulator," 2022 International Conference on Radar,Antenna, Microwave, Electronics, and Telecommunic ations (ICRAMET), Bandung, Indonesia, 2022, pp. 212-218, doi: 10.1109/ICRAMET56917.2022.9991234.

[5]. X. Tian, L. Li, S. Zhao, W. Wang, P. Fu and M. Wang, "Intelligent NAND Flash Memory for In-Situ Block Health Prediction with Machine Learning," 2024 International Conference on Microelectronics (ICM), Doha, Qatar, 2024, pp. 1-5, doi: 10.1109/ICM63406.2024.10815814.

[6]. W. Liu, G. Zhuang, X. Liu, S. Hu, R. He and Y. Wang, "How do we move towards true artificial intelligence," 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Haikou, Hainan, China, 2021, pp. 2156-2158, doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00321

[7]. Y. Gerstorfer, L. Krieg, and M. Hahn-Klimroth, “A Notion of Feature Importance by Decorrelation and Detection of Trends by Random Forest Regression,” arXiv preprint arXiv:2303.01156, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.01156

[8]. B. B. Yuksel and A. Y. Metin, “Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review,” arXiv preprint arXiv:2505.16771, May 2025. [Online]. Available: https://arxiv.org/abs/2505.16771