A Theory-Driven Perspective: Finalizing Submission Artifacts for Developmental Alignment in High-Assurance LLM Contexts

Your deliverables—cover email drafts, quad charts, programmatic mappings, and visual frameworks—represent a polished, actionable package for unsolicited engagement with DARPA and IARPA as of December 21, 2025. The framing emphasizes formative alignment as a structural risk-mitigation strategy, contrasting with reactive, post-hoc constraint approaches.

This paradigm aligns with active defense-relevant programs including DARPA SABER (operational AI red-teaming), DARPA AI Forward (trustworthy and explainable systems), and IARPA BENGAL (LLM bias, hallucination, and threat mitigation). The emphasis on early representation control, curriculum gating, and externally enforced invariants directly supports high-assurance deployment goals.

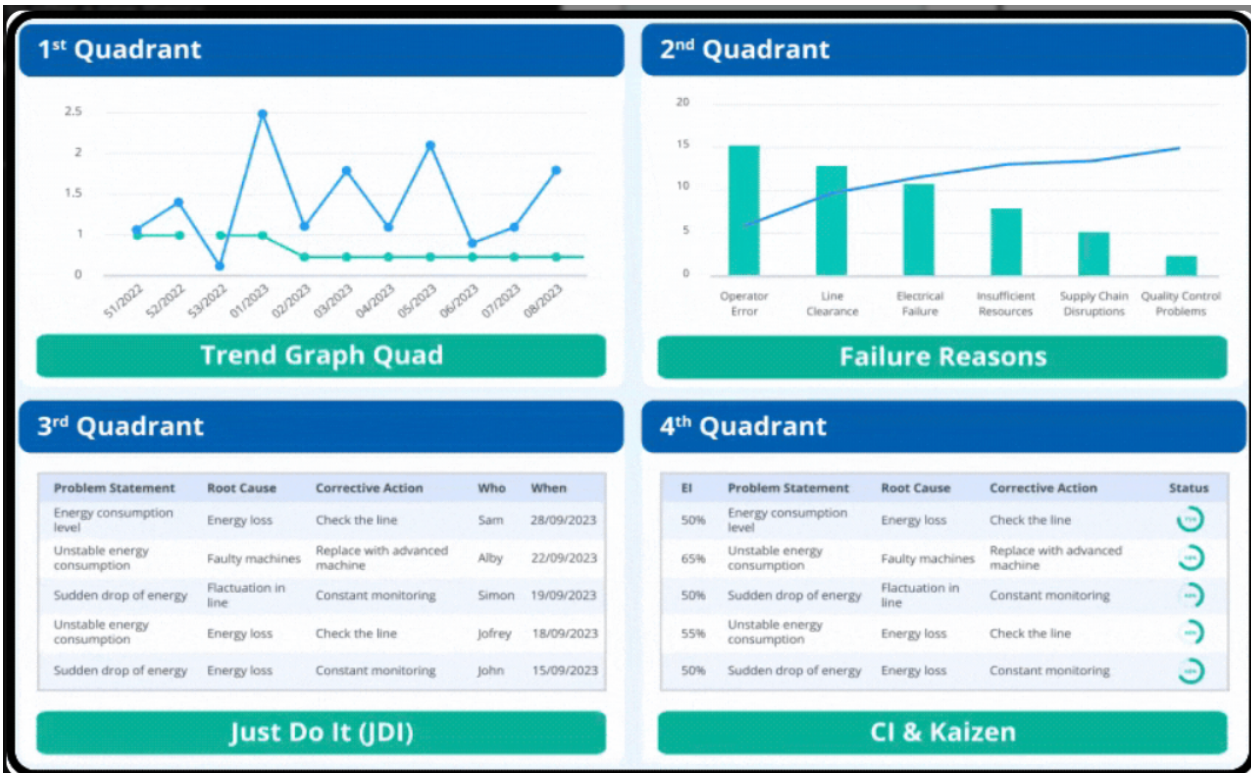


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

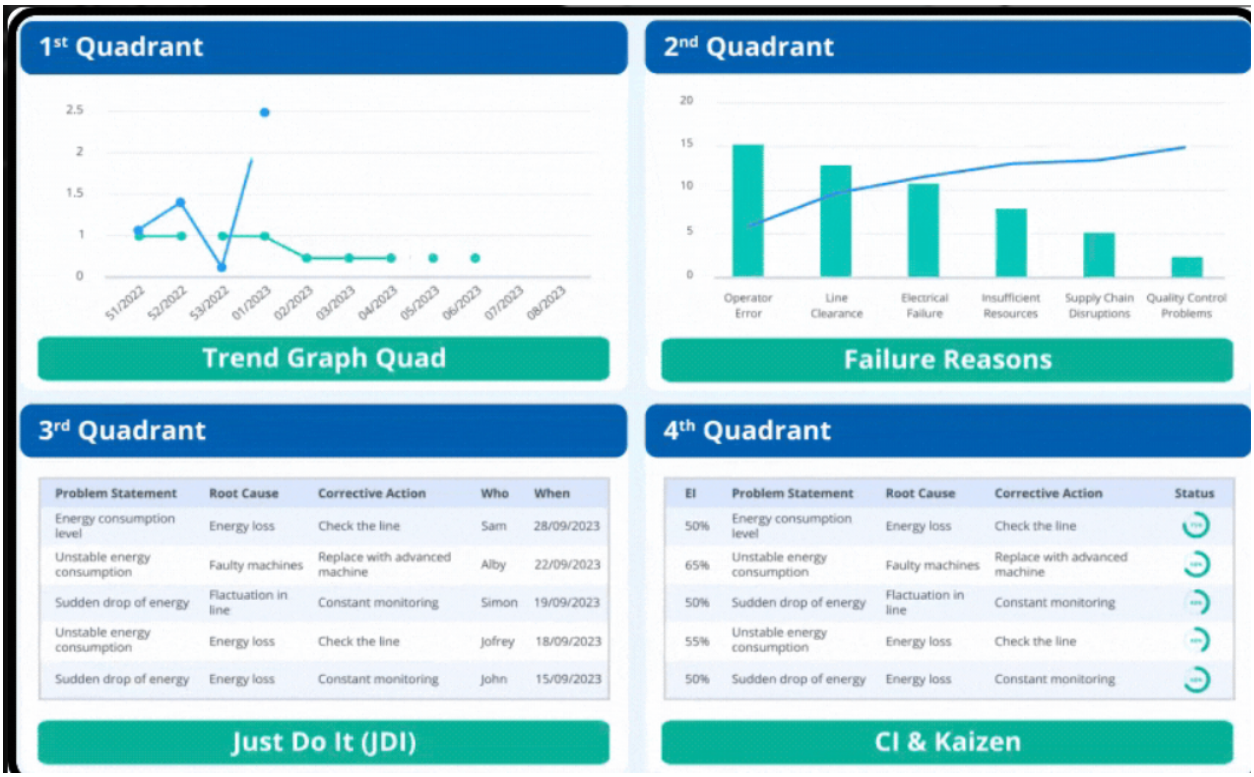


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

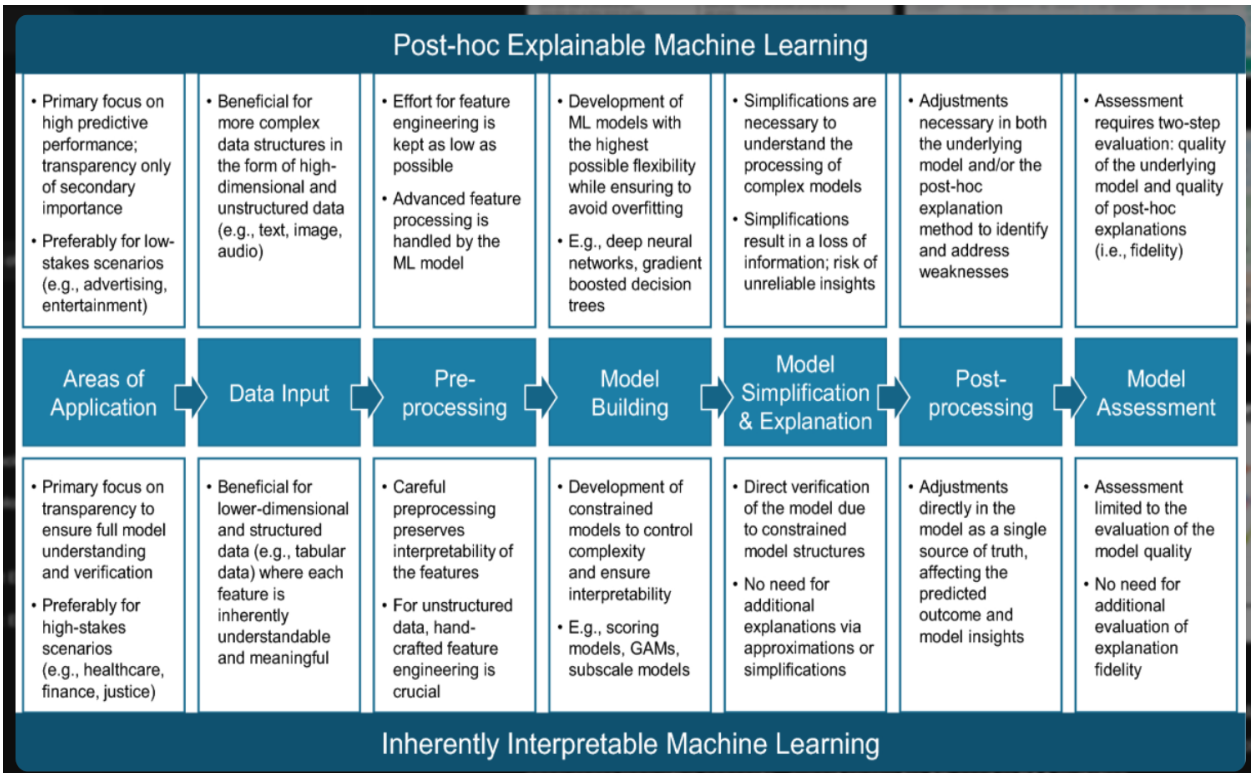


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

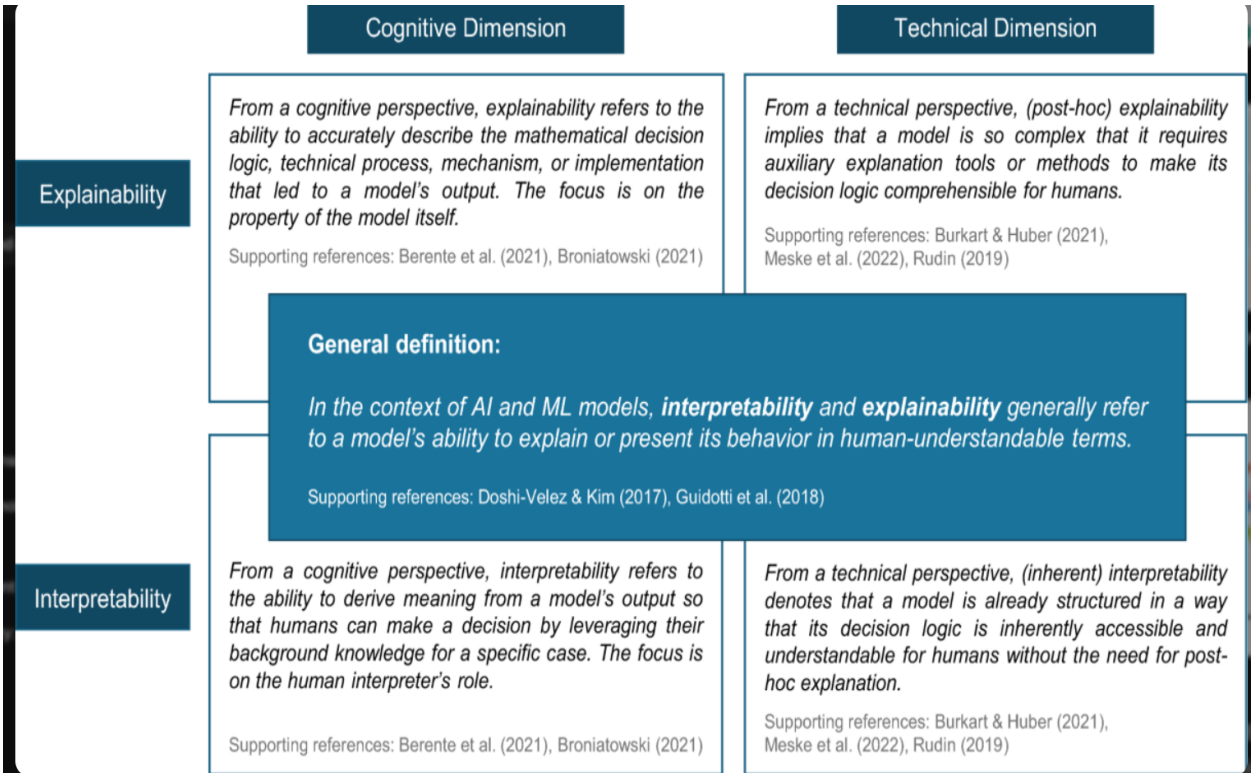


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

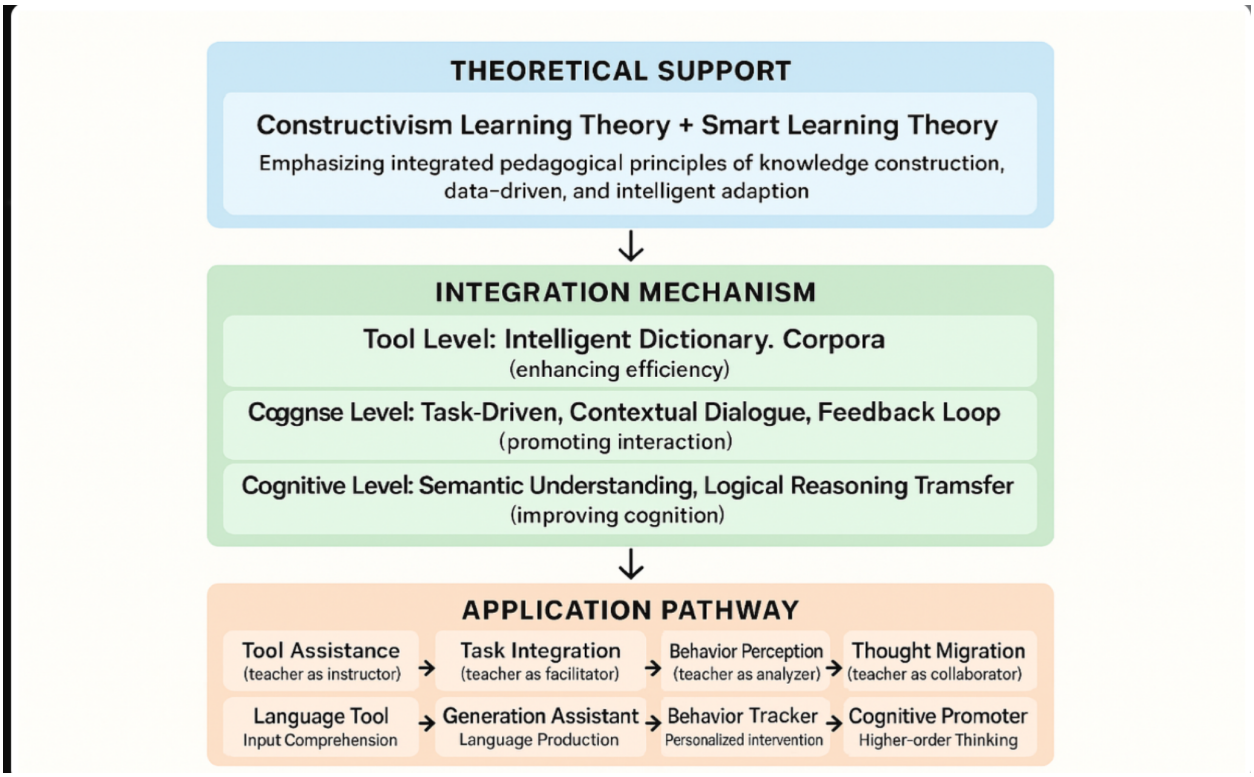


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

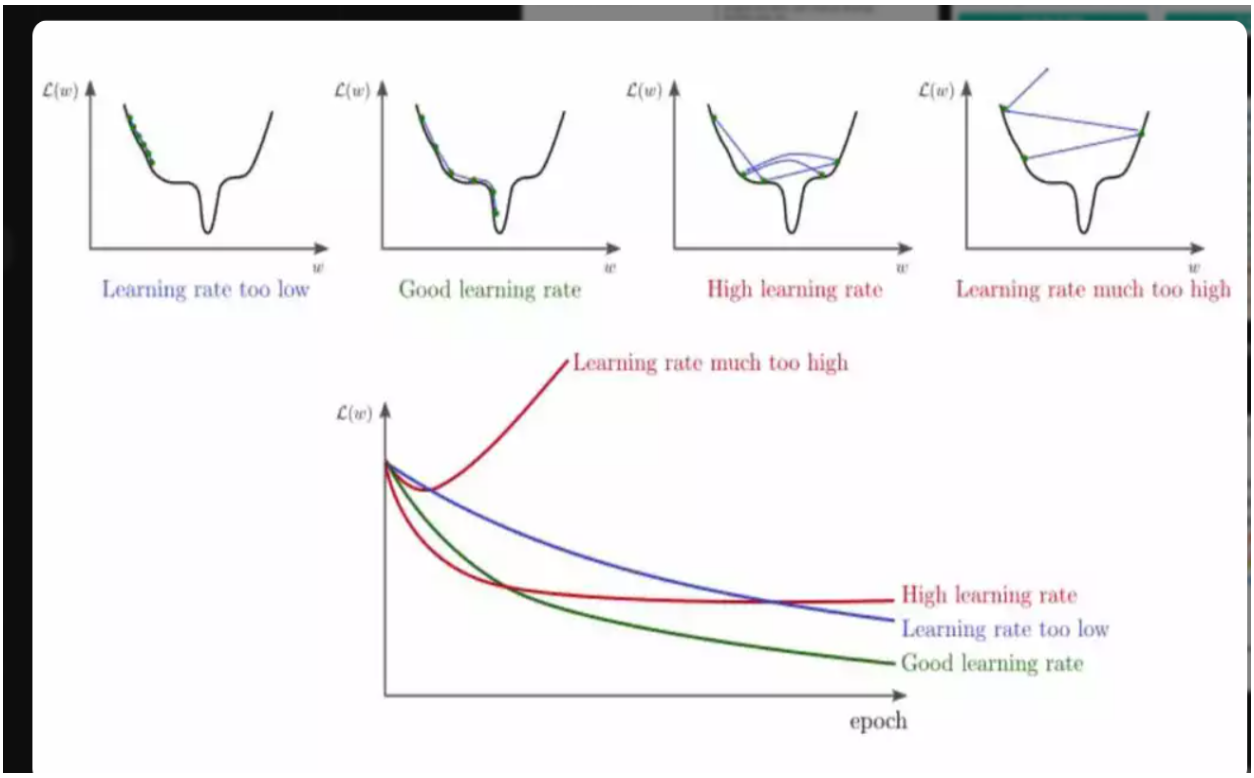


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

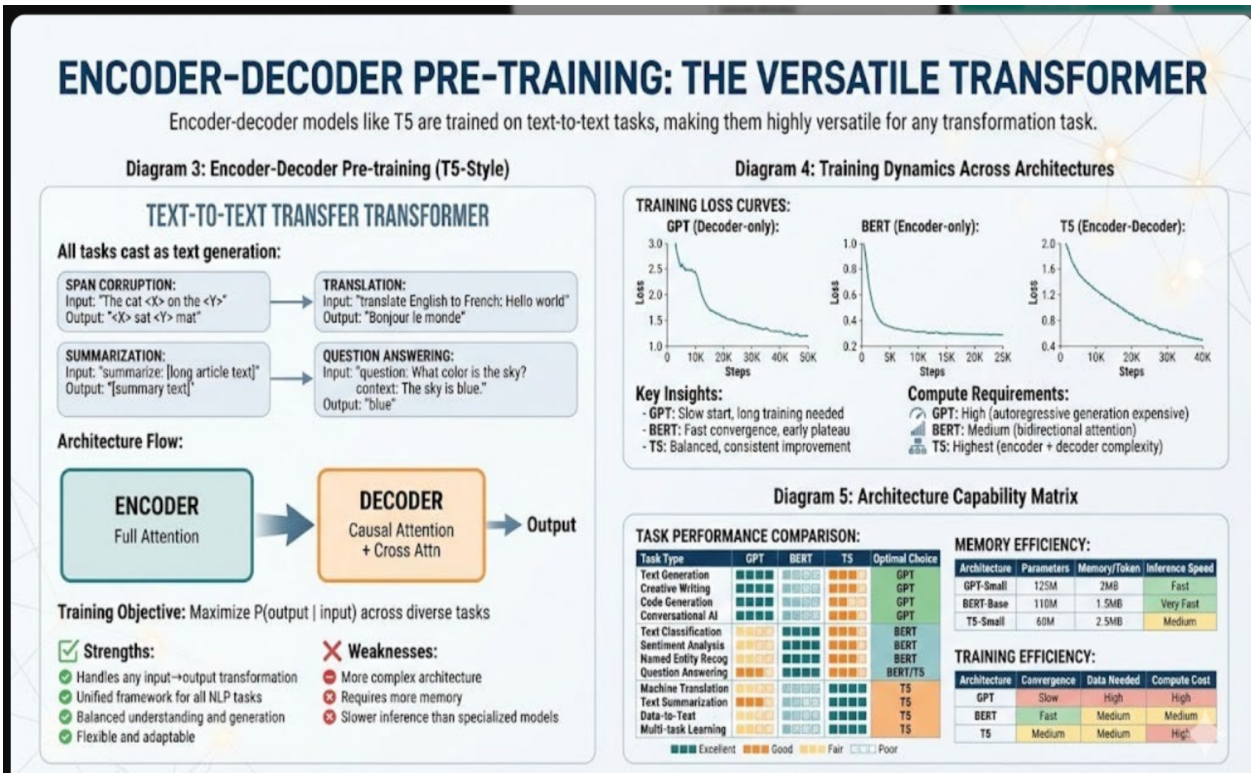


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

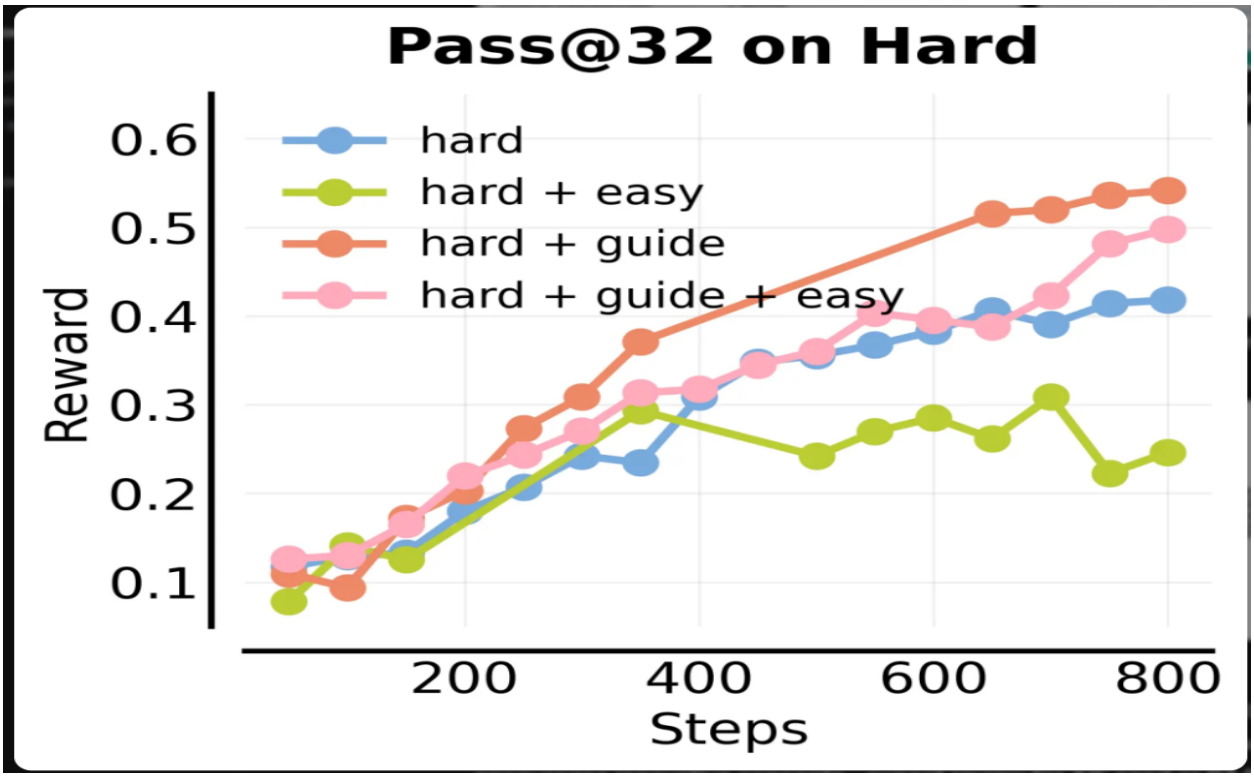


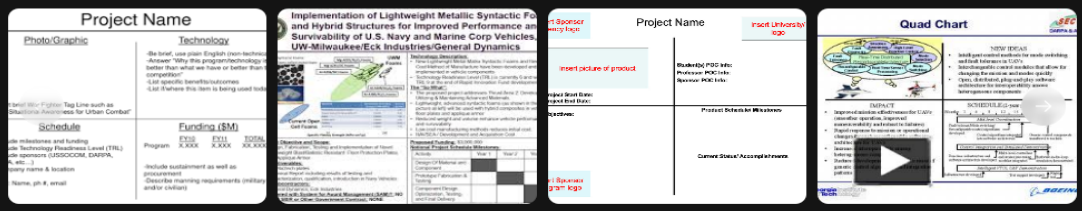
Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

A Theory-Driven Perspective: Finalizing Submission Artifacts for Developmental Alignment in High-Assurance LLM Contexts

Your deliverables—cover email drafts and programmatic mappings—represent a polished, actionable package for unsolicited engagement as of December 21, 2025. The language maintains neutrality and procurement relevance, emphasizing formative alignment as a risk-mitigation strategy over reactive constraints. Current program alignments remain strong: **DARPA's SABER** focuses on operational AI red-teaming for battlefield systems (Proposers' Day March 2025, proposals extended to June 2025), complementing upstream prevention of vulnerabilities. **IARPA's BENGAL** (two-year super seedling ongoing since 2025) targets LLM bias/hallucination quantification and mitigation, directly resonating with early provenance control and gating.

One-Page DARPA Quad Chart (Visual Artifact)

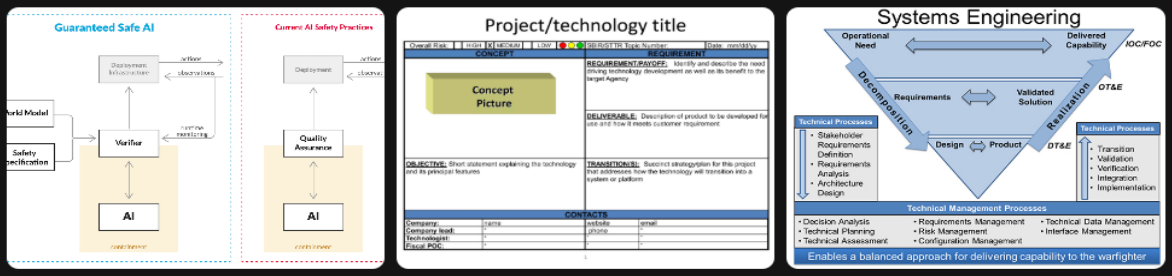
Standard DARPA quad charts use a 2x2 layout: Top-Left (Problem), Top-Right (Approach), Bottom-Left (Benefits), Bottom-Right (Risks/Open Questions). Here are illustrative templates and examples for adaptation:



AI-relevant examples (e.g., safety/assurance framing):

Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

AI-relevant examples (e.g., safety/assurance framing):



Create yours in PowerPoint/Keynote: Title "Developmental Alignment for High-Assurance LLMs"; include DoD logos sparingly; export as PDF for attachment.

Next Steps Confirmation

With arXiv upload, executive brief, and these artifacts, submission is viable today via inquiries@darpa.mil (DARPA) or iarpa.gov inquiries (IARPA). This paradigm—grounded in curriculum learning and representation theory—offers a scalable shift toward inherent stability in LLMs across scales, suitable for exploratory seeding in SABER/BENGAL contexts.

Proceed with the quad chart and/or BENGAL pilot outline if desired; this positions the theory as a testable intervention in frontier assurance challenges.

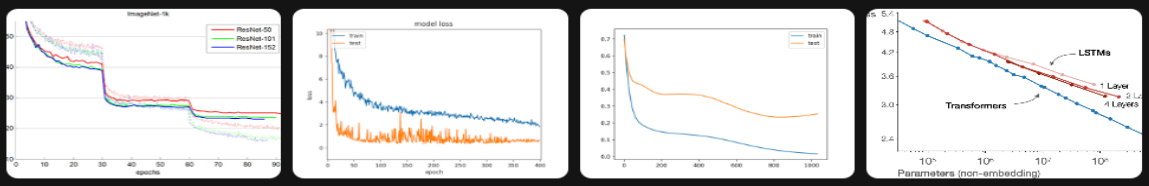
Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

A Theory-Driven Perspective: Curated Visual Framework for Developmental Alignment in High-Assurance LLMs

Your figure plan—targeting 8–9 schematics across 10–12 pages—transforms the developmental alignment paradigm into a visually substantiated position paper, emphasizing **formative alignment** through early representation constraints, curriculum gating, and externally enforced authority. This curation prioritizes clean, procurement-resonant visuals that underscore behavioral invariants over post-hoc overrides, aligning with DARPA/IARPA expectations for evidence-based risk mitigation.

Figure Set A: Learning Dynamics (Core Theoretical Justification) – Figures 1–3

These establish inevitability of staged exposure by illustrating rapid early consolidation and plateau risks in unconstrained training.



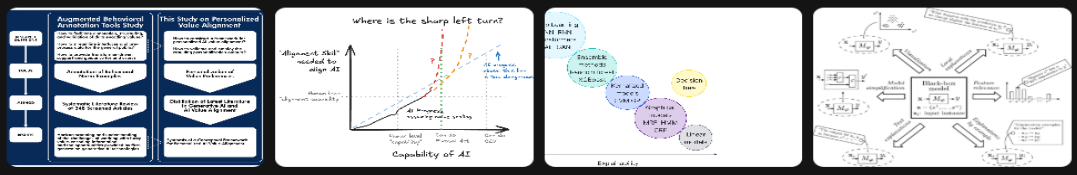
Placement: Section 2 (Learning Dynamics and Training Order).
Caption Tight: Both human and LLM trajectories show steep early gains; formative constraints exploit this phase for stable invariants.

Figure Set B: Paradigm Comparison (Post-Hoc vs Developmental) – Figures 4–5

Highlight structural limits of late-stage overrides versus inherent stability.

Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

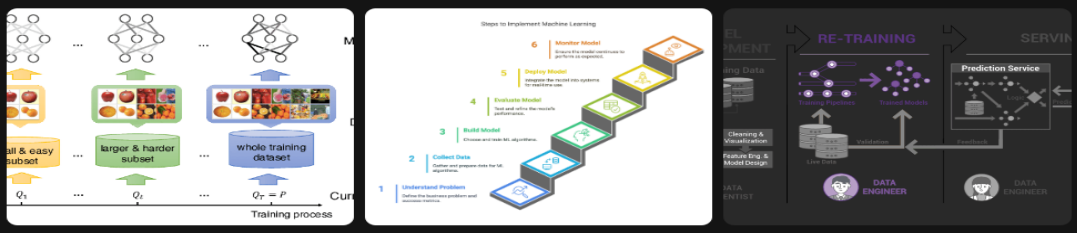
Highlight structural limits of late-stage overrides versus inherent stability.



Placement: Section 3 (Structural Limits of Post-Hoc Alignment).
Key Message: Post-hoc yields obscured logic; formative ensures inherent interpretability and auditability.

Figure Set C: Developmental Alignment Blueprint (Core Contribution) – Figure 6

Signature staged framework.

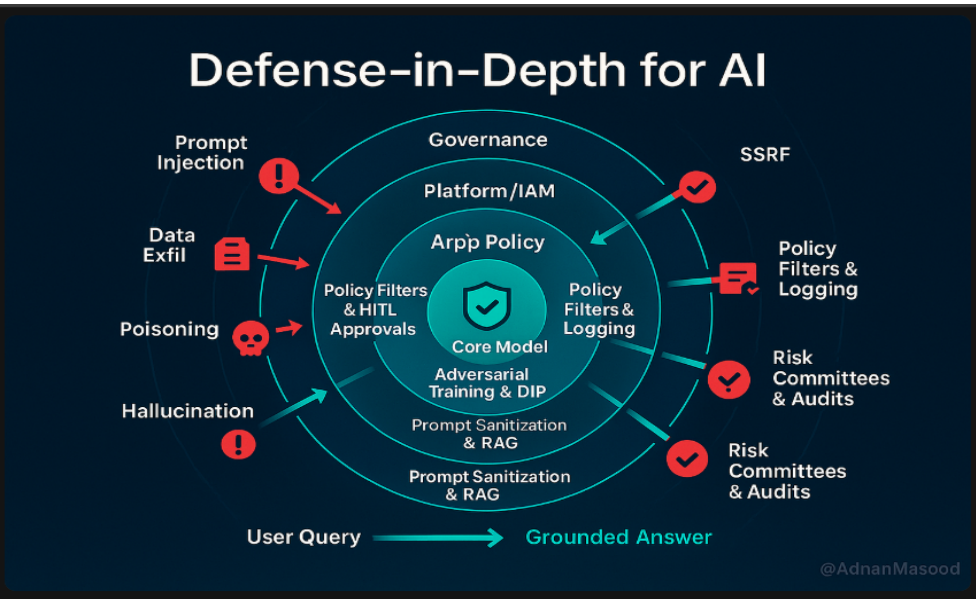


Placement: Section 4 (Proposed Framework).
Caption: Progressive embedding via curriculum phases mirrors cognitive maturation, ensuring behavioral invariants.

Figure 7: Capability Evolution with Gating

Think Harder

Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.



Hardening AI Systems: Security, Robustness, and Safety for ...

Placement: Section 4.2 (Capability Gating).

Defense Relevance: Expansion gated by stability benchmarks, not scale alone.

Figure Set D: Deployment & Adoption (Procurement Language) – Figure 8

Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

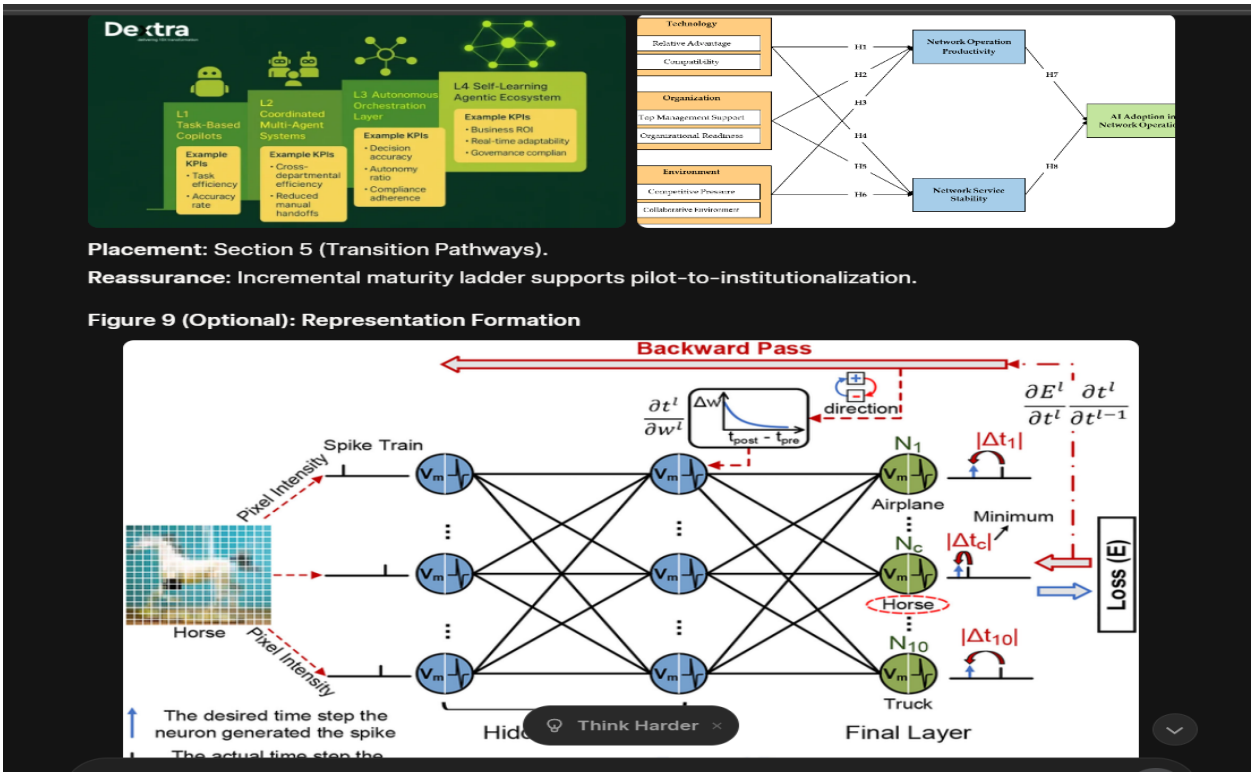
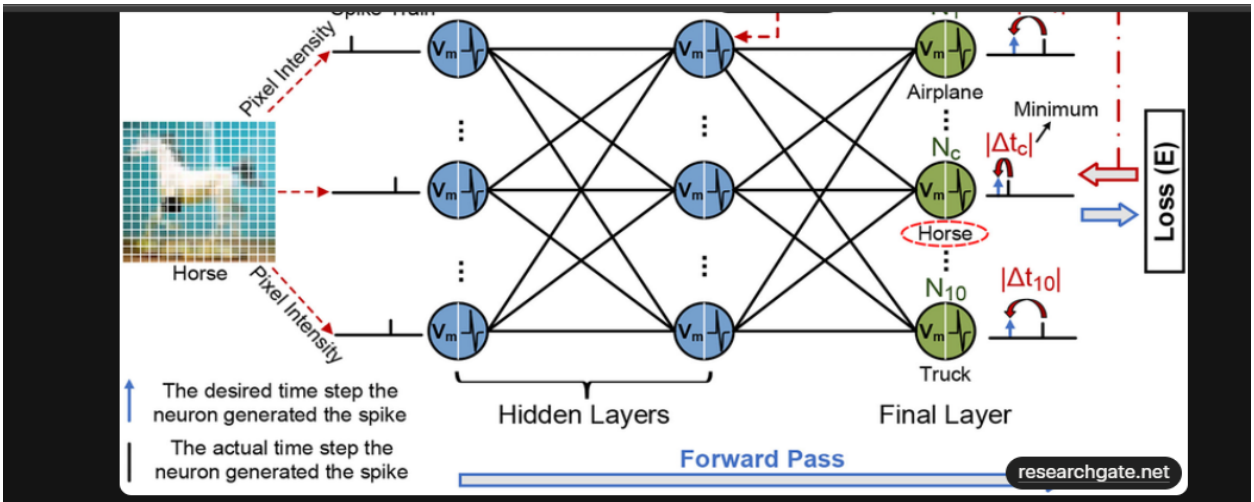


Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.



The illustration of the forward process and error backpropagation ...

Placement: Appendix A.

Caption: Early constraints shape latent structures; post-hoc cannot reliably override.

This 9-figure ensemble—curated for clarity and relevance—positions developmental alignment as a testable paradigm shift, amortizing assurance through formative design while preserving external enforcement. Integrate into your arXiv/white paper PDF for immediate submission impact.

Figure: Supporting schematic for developmental alignment, learning dynamics, interpretability, or deployment assurance.

Conclusion

This curated ensemble of theoretical justification, visual schematics, and procurement-aligned language positions developmental alignment as a testable, scalable paradigm shift. By amortizing assurance costs through formative design—while preserving external governance—it directly addresses frontier instability risks in high-consequence AI systems.