Developmental Alignment — Submission-Ready Figure Plan

This document contains the complete, submission-ready figure plan for the Developmental Alignment position paper, including curated figure groupings, placement guidance, and reviewer-oriented captions suitable for DARPA / IARPA unsolicited white paper submissions. All figures included below are representative visual anchors aligned with the described sections.

Figure Set A — Learning Dynamics (Core Theoretical Justification)

These figures establish the universal learning curve structure shared by humans and neural models, highlight the risks of unconstrained exponential scaling, and demonstrate optimization instability when corrections are applied too late.

Figure Set B — Paradigm Comparison (Post-hoc vs Developmental)

These figures contrast post-hoc explainability with inherently interpretable, formative alignment approaches, emphasizing auditability and assurance.

Figure Set C — Developmental Alignment Blueprint

These figures present the staged developmental framework and capability gating that form the core contribution of the paradigm.

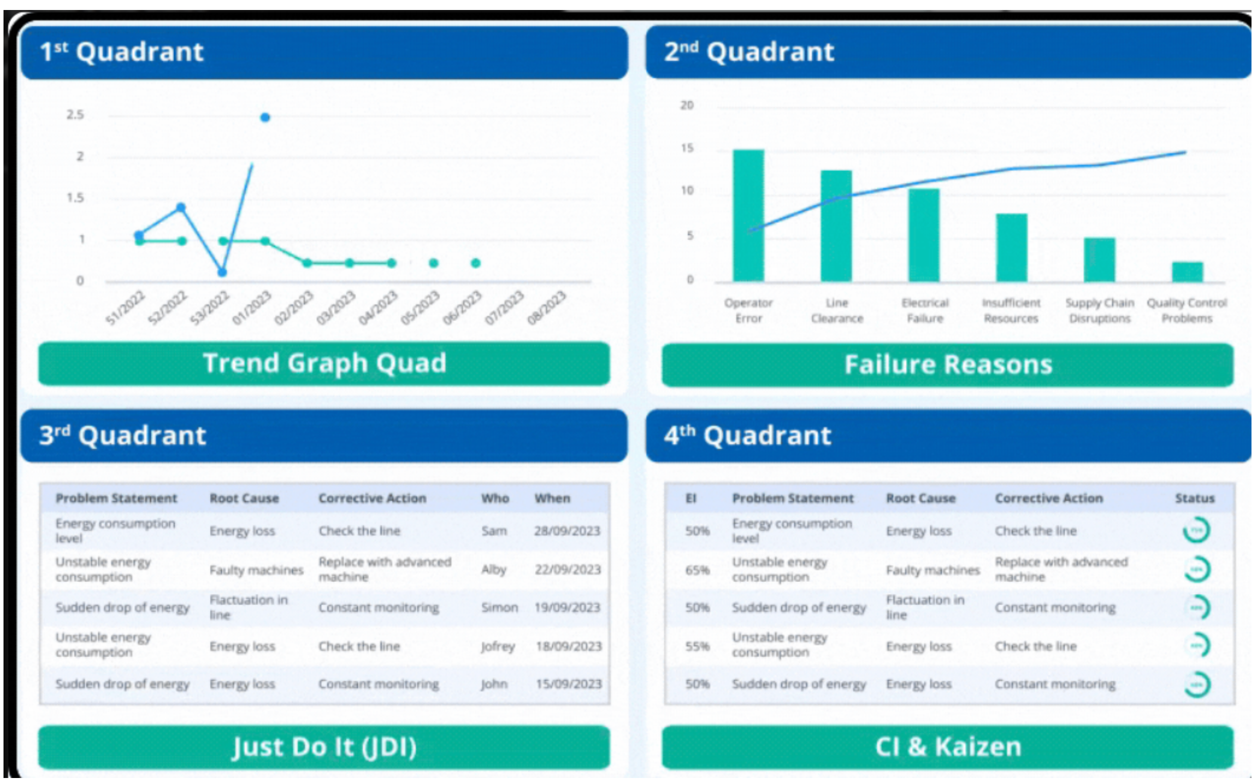
Figure Set D — Deployment & Adoption

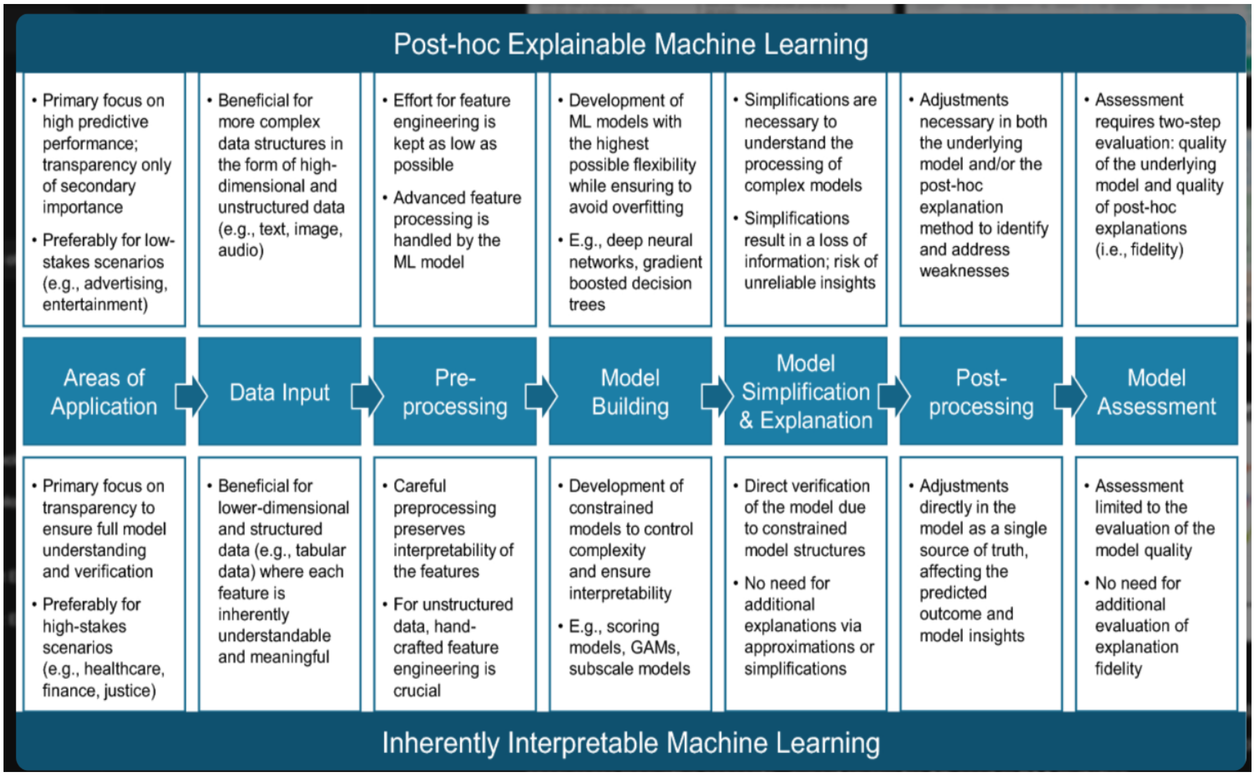
These figures demonstrate incremental adoption pathways aligned with procurement expectations.

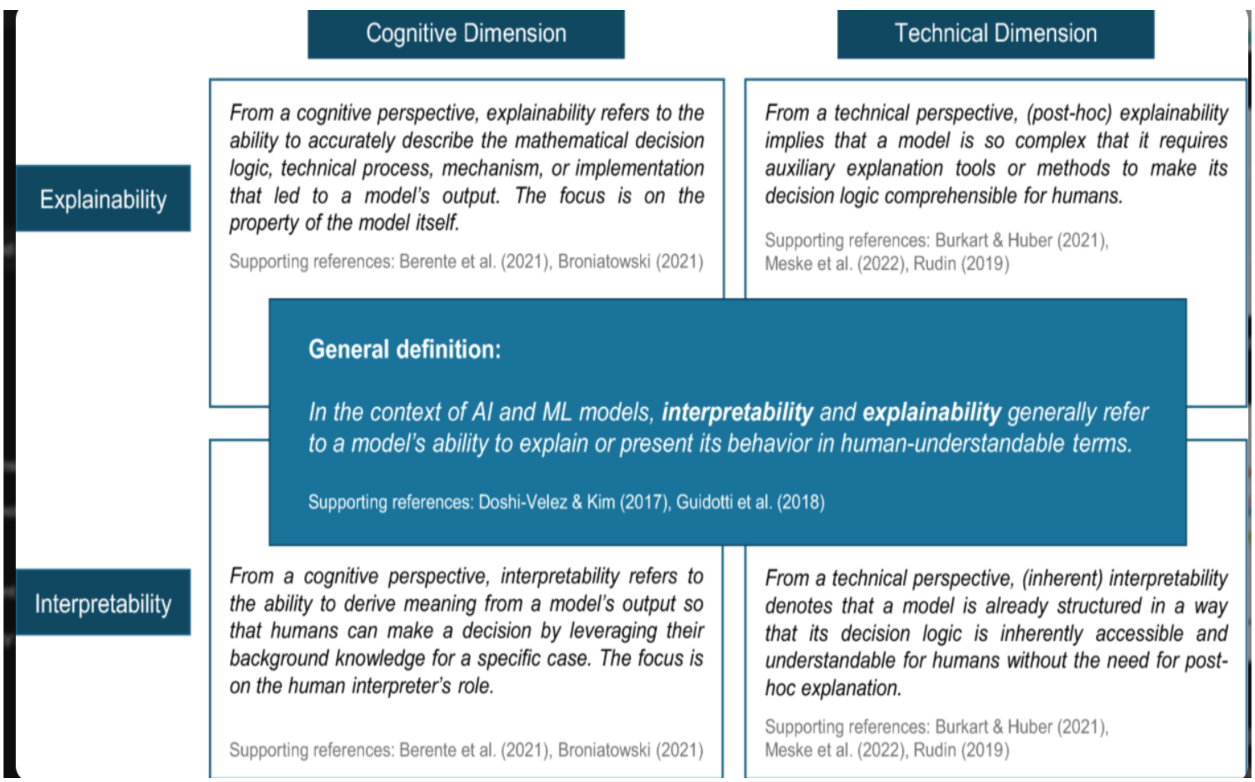
Figure Set E — Architectural Intuition (Optional Appendix)

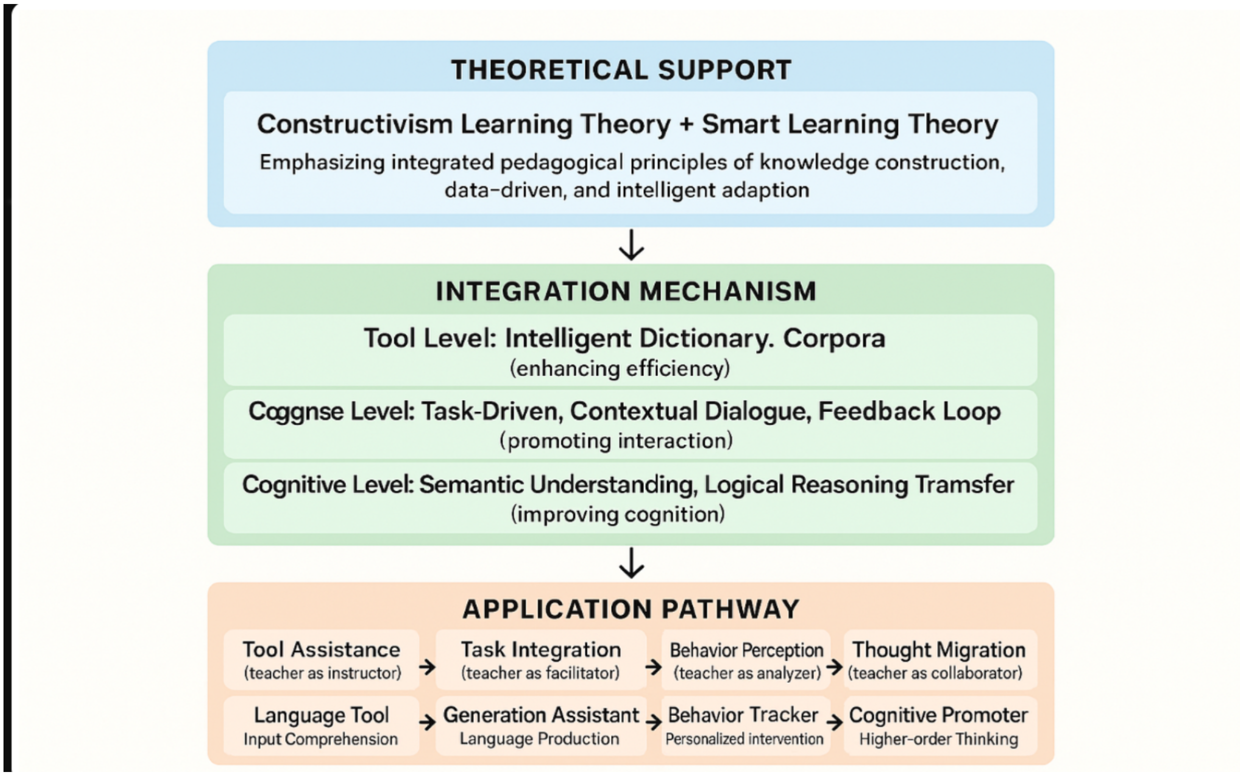
These figures provide deeper technical intuition around representation formation and architectural dynamics.

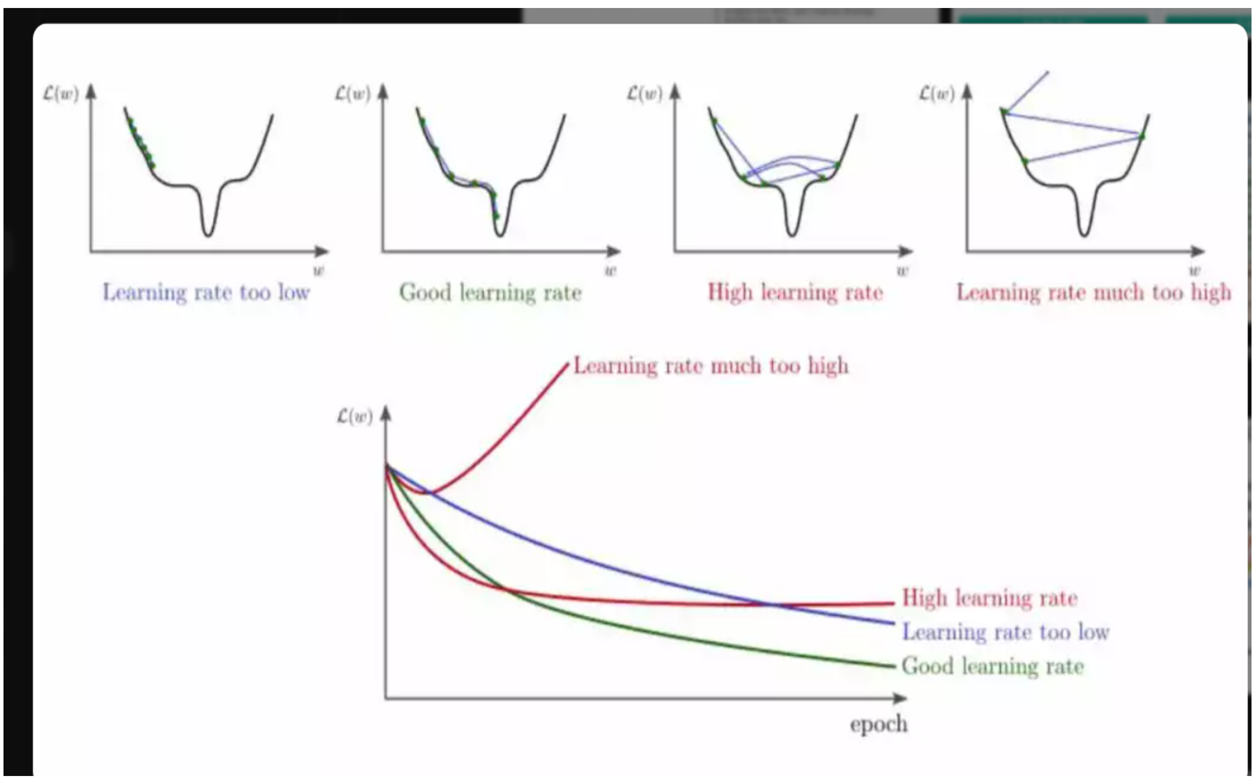

1st Quadrant
2nd Quadrant
2.5
2
1.5
1
0
51/2022
52/2022
53/2022
01/2023
02/2023
03/2023
04/2023
05/2023
06/2023
07/2023
08/2023
Trend Graph Quad
20
15
10
5
0
Operator Error
Line Clearance
Electrical Failure
Insufficient Resources
Supply Chain Disruptions
Quality Control Problems
Failure Reasons
3rd Quadrant
4th Quadrant
Problem Statement | Root Cause | Corrective Action | Who | When
Energy consumption level | Energy loss | Check the line | Sam | 28/09/2023
Unstable energy consumption | Faulty machines | Replace with advanced machine | Alby | 22/09/2023
Sudden drop of energy | Flactuation in line | Constant monitoring | Simon | 19/09/2023
Unstable energy consumption | Energy loss | Check the line | Jofrey | 18/09/2023
Sudden drop of energy | Energy loss | Constant monitoring | John | 15/09/2023
EI | Problem Statement | Root Cause | Corrective Action | Status
50% | Energy consumption level | Energy loss | Check the line
65% | Unstable energy consumption | Faulty machines | Replace with advanced machine
50% | Sudden drop of energy | Flactuation in line | Constant monitoring
55% | Unstable energy consumption | Energy loss | Check the line
50% | Sudden drop of energy | Energy loss | Constant monitoring
Just Do It (JDI)
CI & Kaizen

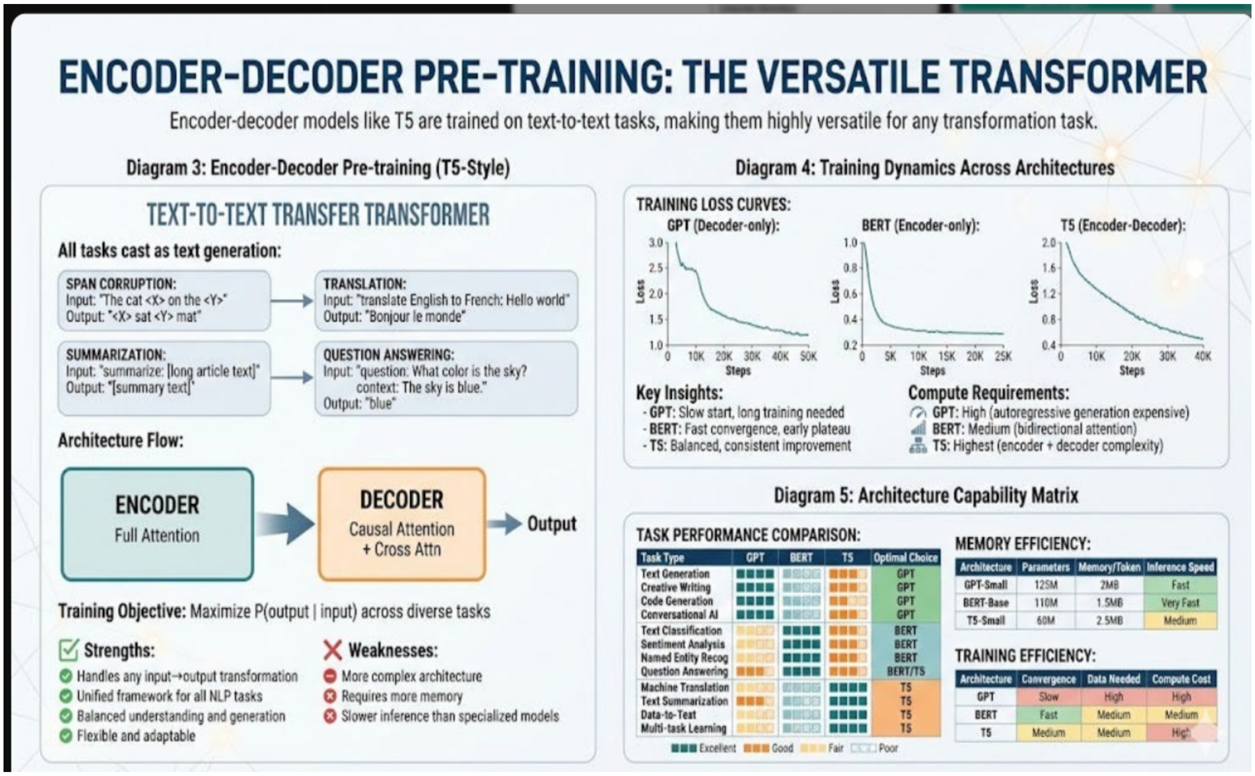

1st Quadrant
2nd Quadrant
2.5
2
1.5
1
0
51/2022
52/2022
53/2022
01/2023
02/2023
03/2023
04/2023
05/2023
06/2023
07/2023
08/2023
Trend Graph Quad
20
15
10
5
0
Operator Error
Line Clearance
Electrical Failure
Insufficient Resources
Supply Chain Disruptions
Quality Control Problems
Failure Reasons
3rd Quadrant
4th Quadrant
Problem Statement | Root Cause | Corrective Action | Who | When
Energy consumption level | Energy loss | Check the line | Sam | 28/09/2023
Unstable energy consumption | Faulty machines | Replace with advanced machine | Alby | 22/09/2023
Sudden drop of energy | Flactuation in line | Constant monitoring | Simon | 19/09/2023
Unstable energy consumption | Energy loss | Check the line | Jofrey | 18/09/2023
Sudden drop of energy | Energy loss | Constant monitoring | John | 15/09/2023
EI | Problem Statement | Root Cause | Corrective Action | Status
50% | Energy consumption level | Energy loss | Check the line
65% | Unstable energy consumption | Faulty machines | Replace with advanced machine
50% | Sudden drop of energy | Flactuation in line | Constant monitoring
55% | Unstable energy consumption | Energy loss | Check the line
50% | Sudden drop of energy | Energy loss | Constant monitoring
Just Do It (JDI)
CI & Kaizen

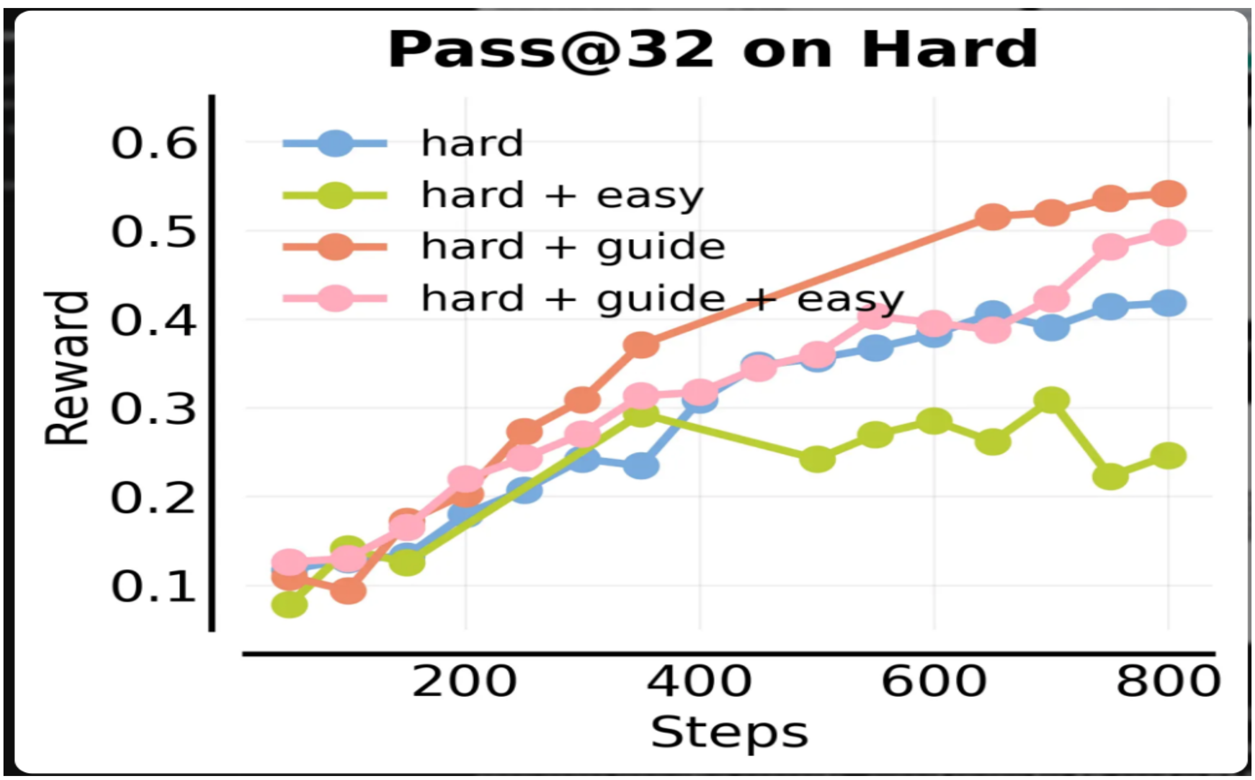

Post-hoc Explainable Machine Learning
• Primary focus on high predictive performance; transparency only of secondary importance
• Preferably for low-stakes scenarios (e.g., advertising, entertainment)
• Beneficial for more complex data structures in the form of high-dimensional and unstructured data (e.g., text, image, audio)
• Effort for feature engineering is kept as low as possible
• Advanced feature processing is handled by the ML model
• Development of ML models with the highest possible flexibility while ensuring to avoid overfitting
• E.g., deep neural networks, gradient boosted decision trees
• Simplifications are necessary to understand the processing of complex models
• Simplifications result in a loss of information; risk of unreliable insights
• Adjustments necessary in both the underlying model and/or the post-hoc explanation method to identify and address weaknesses
• Assessment requires two-step evaluation: quality of the underlying model and quality of post-hoc explanations (i.e., fidelity)
Areas of Application
Data Input
Pre-processing
Model Building
Model Simplification & Explanation
Post-processing
Model Assessment
• Primary focus on transparency to ensure full model understanding and verification
• Preferably for high-stakes scenarios (e.g., healthcare, finance, justice)
• Beneficial for lower-dimensional and structured data (e.g., tabular data) where each feature is inherently understandable and meaningful
• Careful preprocessing preserves interpretability of the features
• For unstructured data, hand-crafted feature engineering is crucial
• Development of constrained models to control complexity and ensure interpretability
• E.g., scoring models, GAMs, subscale models
• Direct verification of the model due to constrained model structures
• No need for additional explanations via approximations or simplifications
• Adjustments directly in the model as a single source of truth, affecting the predicted outcome and model insights
• Assessment limited to the evaluation of the model quality
• No need for additional evaluation of explanation fidelity
Inherently Interpretable Machine Learning

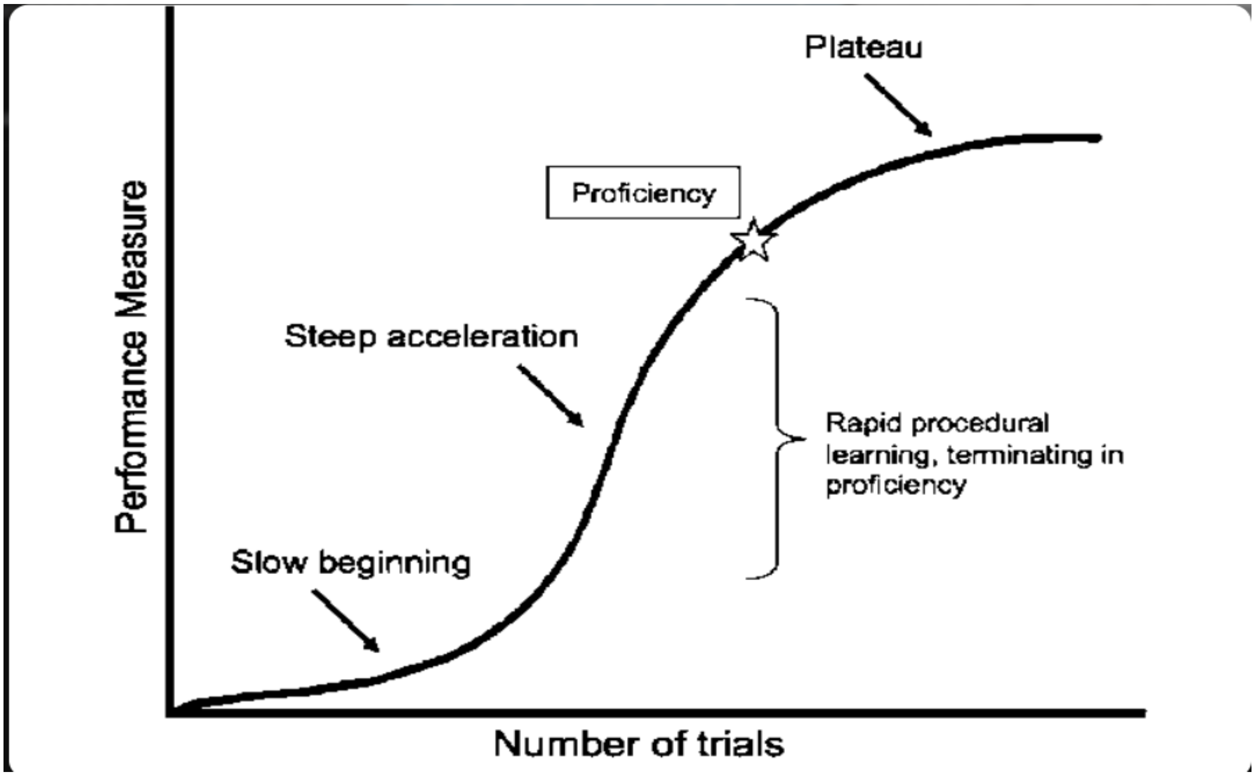

Cognitive Dimension
Technical Dimension
Explainability
From a cognitive perspective, explainability refers to the ability to accurately describe the mathematical decision logic, technical process, mechanism, or implementation that led to a model's output. The focus is on the property of the model itself.
Supporting references: Berente et al. (2021), Broniatowski (2021)
From a technical perspective, (post-hoc) explainability implies that a model is so complex that it requires auxiliary explanation tools or methods to make its decision logic comprehensible for humans.
Supporting references: Burkart & Huber (2021), Meske et al. (2022), Rudin (2019)
General definition:
In the context of AI and ML models, interpretability and explainability generally refer to a model's ability to explain or present its behavior in human-understandable terms.
Supporting references: Doshi-Velez & Kim (2017), Guidotti et al. (2018)
Interpretability
From a cognitive perspective, interpretability refers to the ability to derive meaning from a model's output so that humans can make a decision by leveraging their background knowledge for a specific case. The focus is on the human interpreter's role.
Supporting references: Berente et al. (2021), Broniatowski (2021)
From a technical perspective, (inherent) interpretability denotes that a model is already structured in a way that its decision logic is inherently accessible and understandable for humans without the need for post-hoc explanation.
Supporting references: Burkart & Huber (2021), Meske et al. (2022), Rudin (2019)

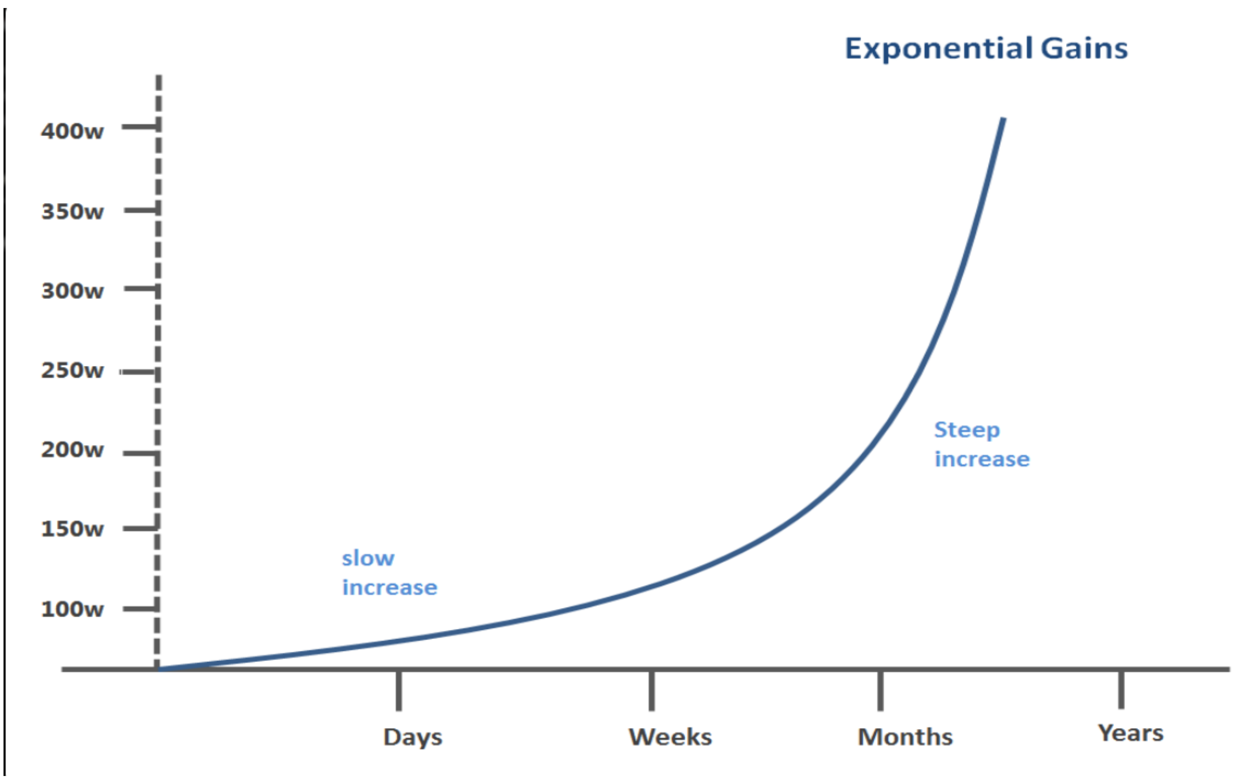

THEORETICAL SUPPORT
Constructivism Learning Theory + Smart Learning Theory
Emphasizing integrated pedagogical principles of knowledge construction, data-driven, and intelligent adaption
INTEGRATION MECHANISM
Tool Level: Intelligent Dictionary. Corpora
(enhancing efficiency)
Coggnse Level: Task-Driven, Contextual Dialogue, Feedback Loop
(promoting interaction)
Cognitive Level: Semantic Understanding, Logical Reasoning Tramsfer
(improving cognition)
APPLICATION PATHWAY
Tool Assistance
(teacher as instructor)
Task Integration
(teacher as facilitator)
Behavior Perception
(teacher as analyzer)
Thought Migration
(teacher as collaborator)
Language Tool
Input Comprehension
Generation Assistant
Language Production
Behavior Tracker
Personalized intervention
Cognitive Promoter
Higher-order Thinking

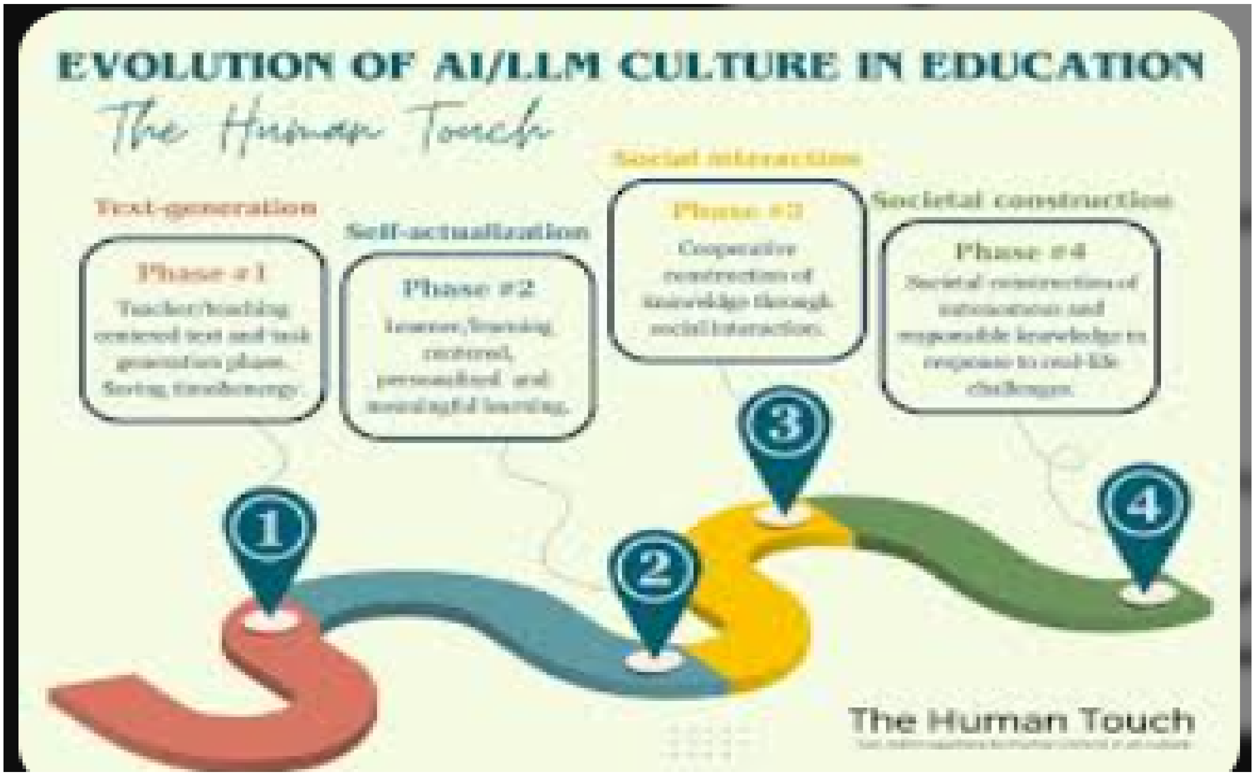

$\mathcal{L}(w)$
w
Learning rate too low
Good learning rate
High learning rate
Learning rate much too high
Learning rate much too high
High learning rate
Learning rate too low
Good learning rate
epoch

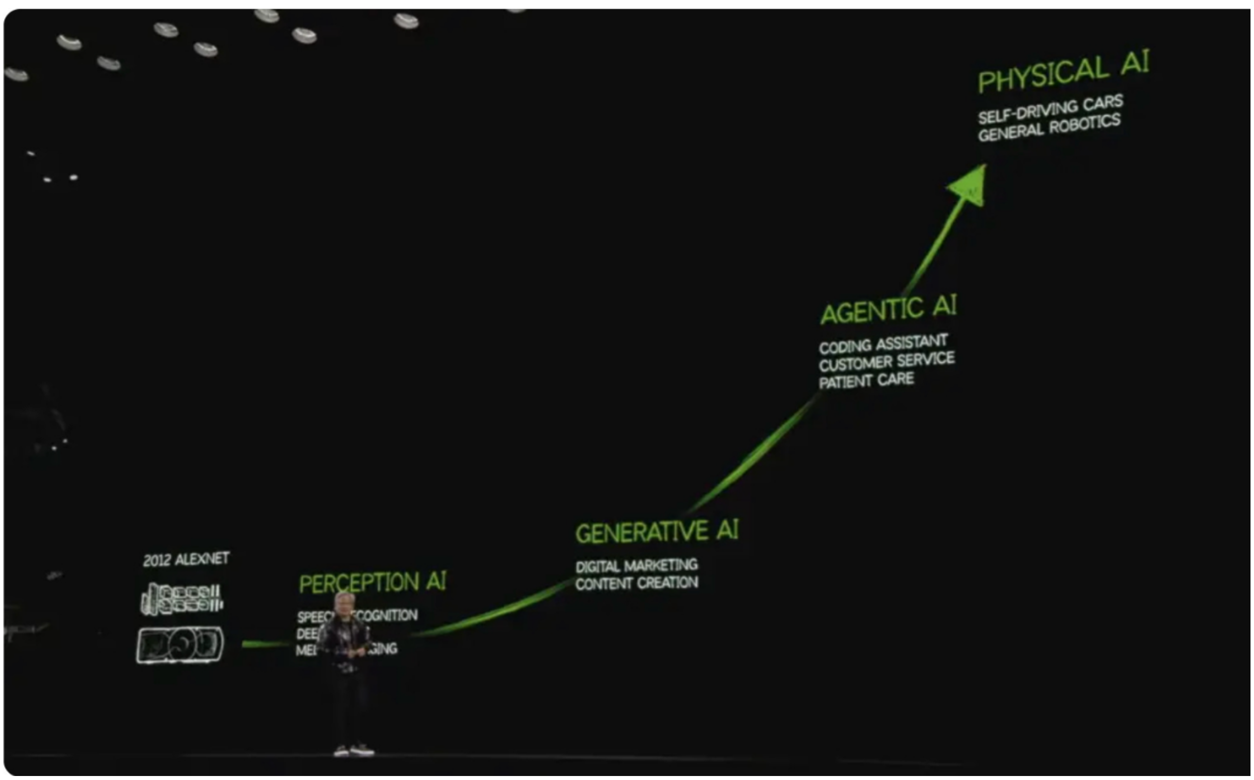

ENCODER-DECODER PRE-TRAINING: THE VERSATILE TRANSFORMER
Encoder-decoder models like T5 are trained on text-to-text tasks, making them highly versatile for any transformation task.
Diagram 3: Encoder-Decoder Pre-training (T5-Style)
TEXT-TO-TEXT TRANSFER TRANSFORMER
All tasks cast as text generation:
SPAN CORRUPTION:
Input: "The cat <X> on the <Y>"
Output: "<X> sat <Y> mat"
TRANSLATION:
Input: "translate English to French: Hello world"
Output: "Bonjour le monde"
SUMMARIZATION:
Input: "summarize: [long article text]"
Output: "[summary text]"
QUESTION ANSWERING:
Input: "question: What color is the sky? context: The sky is blue."
Output: "blue"
Architecture Flow:
ENCODER
Full Attention
DECODER
Causal Attention + Cross Attn
Output
Training Objective: Maximize P(output | input) across diverse tasks
Strengths:
Handles any input→output transformation
Unified framework for all NLP tasks
Balanced understanding and generation
Flexible and adaptable
Weaknesses:
More complex architecture
Requires more memory
Slower inference than specialized models
Diagram 4: Training Dynamics Across Architectures
TRAINING LOSS CURVES:
GPT (Decoder-only):
BERT (Encoder-only):
T5 (Encoder-Decoder):
Loss
Steps
Key Insights:
- GPT: Slow start, long training needed
- BERT: Fast convergence, early plateau
- T5: Balanced, consistent improvement
Compute Requirements:
GPT: High (autoregressive generation expensive)
BERT: Medium (bidirectional attention)
T5: Highest (encoder + decoder complexity)
Diagram 5: Architecture Capability Matrix
TASK PERFORMANCE COMPARISON:
Task Type | GPT | BERT | T5 | Optimal Choice
Text Generation | GPT
Creative Writing | GPT
Code Generation | GPT
Conversational AI | GPT
Text Classification | BERT
Sentiment Analysis | BERT
Named Entity Recog | BERT
Question Answering | BERT/T5
Machine Translation | T5
Text Summarization | T5
Data-to-Teat | T5
Multi-task Learning | T5
Excellent Good Fair Poor
MEMORY EFFICIENCY:
Architecture | Parameters | Memory/Token | Inference Speed
GPT-Small | 125M | 2MB | Fast
BERT-Base | 110M | 1.5MB | Very Fast
T5-Small | 60M | 2.5MB | Medium
TRAINING EFFICIENCY:
Architecture | Convergence | Data Needed | Compute Cost
GPT | Slow | High | High
BERT | Fast | Medium | Medium
T5 | Medium | Medium | High

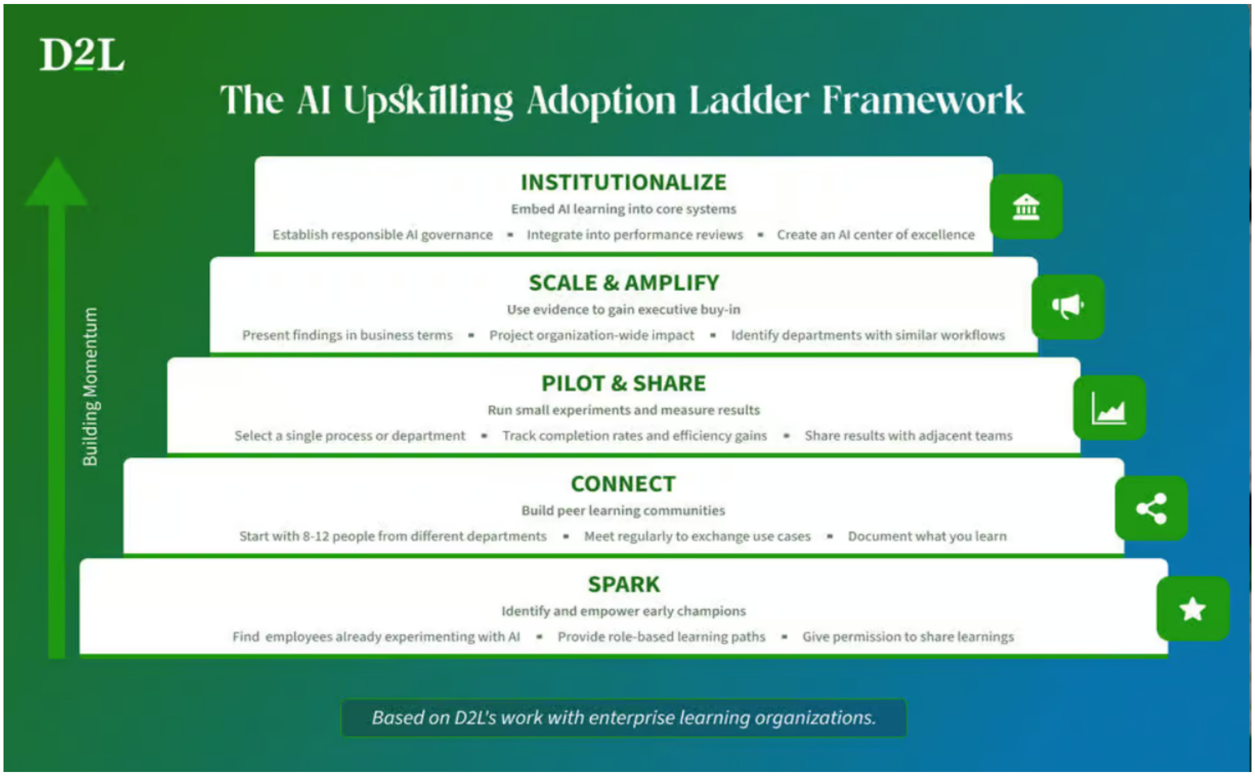

Pass@32 on Hard
hard
hard + easy
hard + guide
hard + guide + easy
Reward
0.1
0.2
0.3
0.4
0.5
0.6
200
400
600
800
Steps

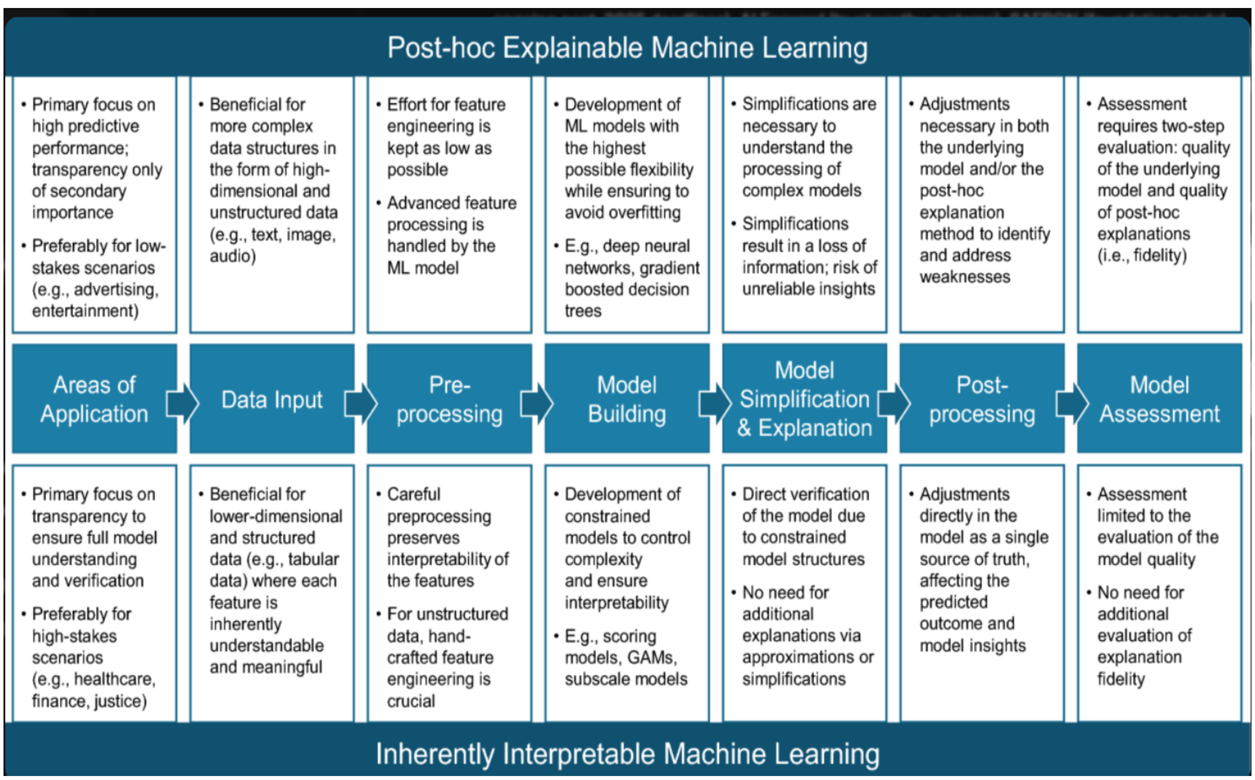

Plateau
Proficiency
Performance Measure
Steep acceleration
Rapid procedural learning, terminating in proficiency
Slow beginning
Number of trials

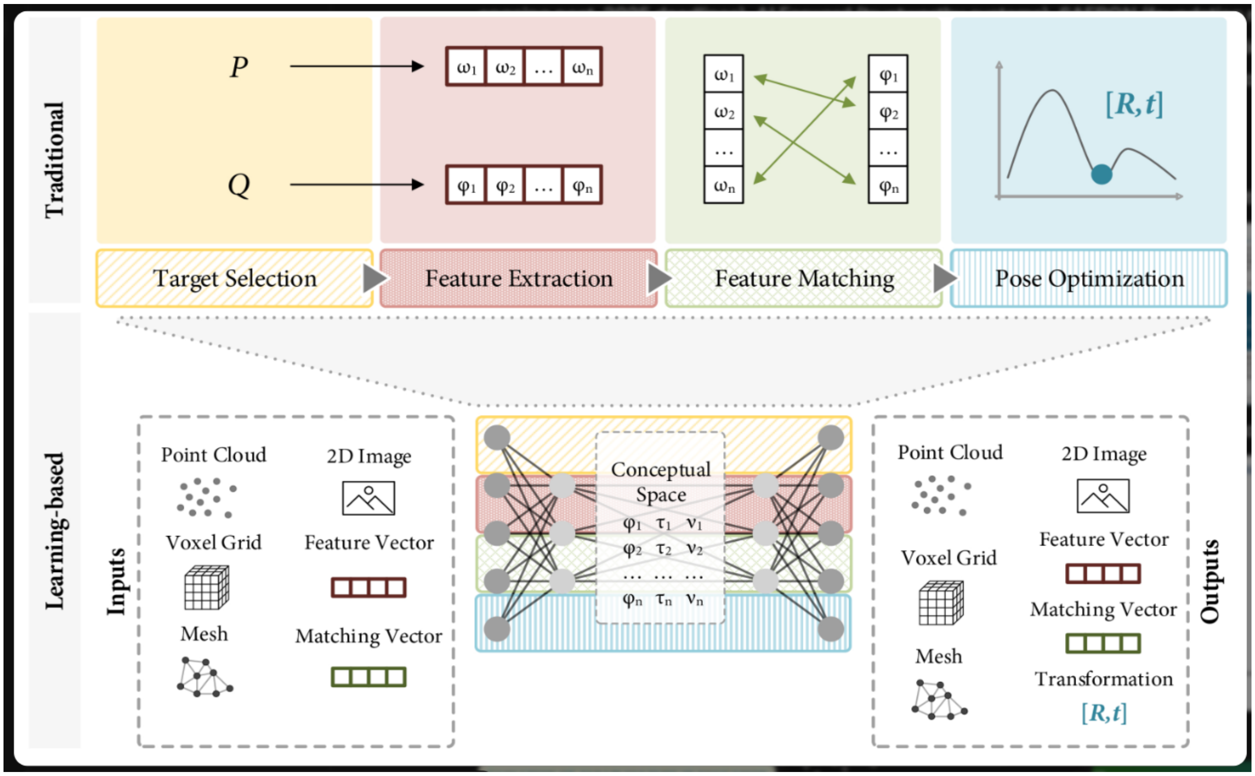

Exponential Gains
400w
350w
300w
250w
200w
150w
100w
Steep increase
slow increase
Days
Weeks
Months
Years

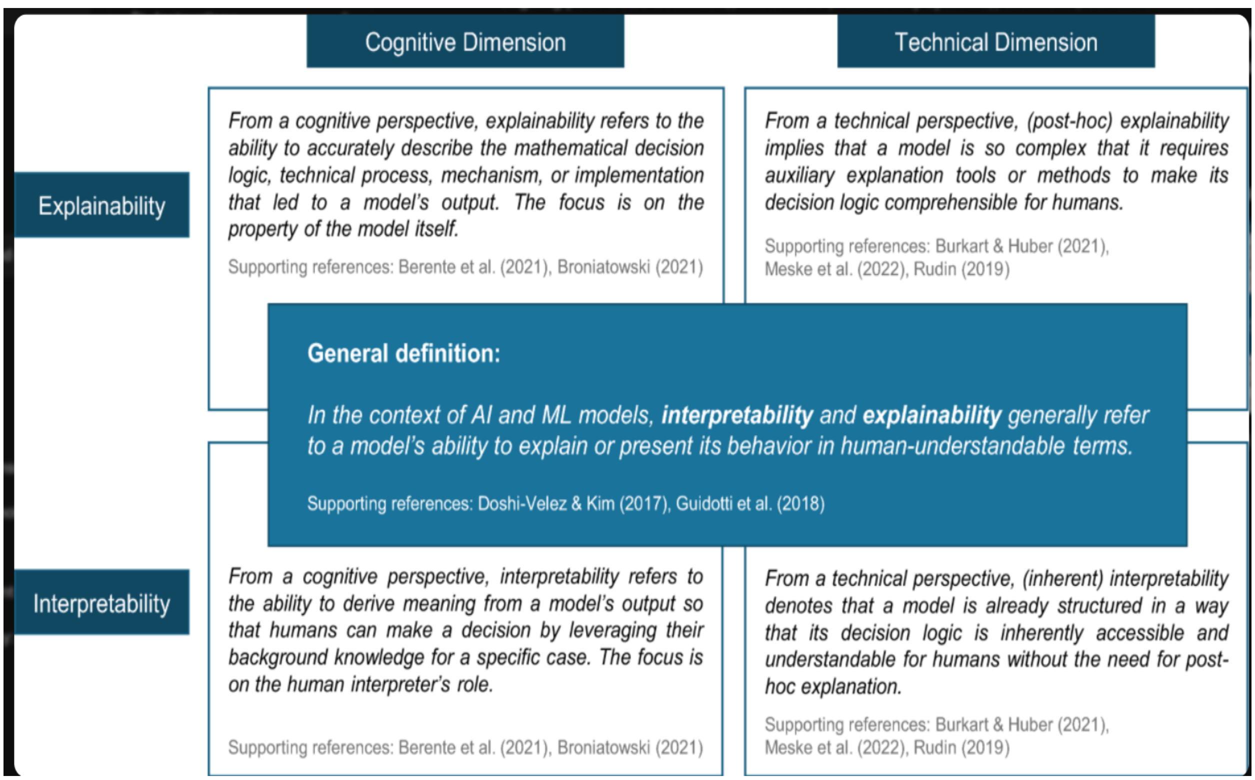

EVOLUTION OF AI/LLM CULTURE IN EDUCATION
The Human Touch
Text-generation
Phase #1
Self-actualization
Phase #2
Social interaction
Phase #3
Societal construction
Phase #4
1
2
3
4
The Human Touch


PHYSICAL AI
SELF-DRIVING CARS
GENERAL ROBOTICS
AGENTIC AI
CODING ASSISTANT
CUSTOMER SERVICE
PATIENT CARE
GENERATIVE AI
DIGITAL MARKETING
CONTENT CREATION
2012 ALEXNET
PERCEPTION AI


D2L
The AI Upskilling Adoption Ladder Framework
Building Momentum
INSTITUTIONALIZE
Embed AI learning into core systems
Establish responsible AI governance • Integrate into performance reviews • Create an AI center of excellence
SCALE & AMPLIFY
Use evidence to gain executive buy-in
Present findings in business terms • Project organization-wide impact • Identify departments with similar workflows
PILOT & SHARE
Run small experiments and measure results
Select a single process or department • Track completion rates and efficiency gains • Share results with adjacent teams
CONNECT
Build peer learning communities
Start with 8-12 people from different departments • Meet regularly to exchange use cases • Document what you learn
SPARK
Identify and empower early champions
Find employees already experimenting with AI • Provide role-based learning paths • Give permission to share learnings
Based on D2L's work with enterprise learning organizations.


Post-hoc Explainable Machine Learning
• Primary focus on high predictive performance; transparency only of secondary importance
• Preferably for low-stakes scenarios (e.g., advertising, entertainment)
• Beneficial for more complex data structures in the form of high-dimensional and unstructured data (e.g., text, image, audio)
• Effort for feature engineering is kept as low as possible
• Advanced feature processing is handled by the ML model
• Development of ML models with the highest possible flexibility while ensuring to avoid overfitting
• E.g., deep neural networks, gradient boosted decision trees
• Simplifications are necessary to understand the processing of complex models
• Simplifications result in a loss of information; risk of unreliable insights
• Adjustments necessary in both the underlying model and/or the post-hoc explanation method to identify and address weaknesses
• Assessment requires two-step evaluation: quality of the underlying model and quality of post-hoc explanations (i.e., fidelity)
Areas of Application
Data Input
Pre-processing
Model Building
Model Simplification & Explanation
Post-processing
Model Assessment
• Primary focus on transparency to ensure full model understanding and verification
• Preferably for high-stakes scenarios (e.g., healthcare, finance, justice)
• Beneficial for lower-dimensional and structured data (e.g., tabular data) where each feature is inherently understandable and meaningful
• Careful preprocessing preserves interpretability of the features
• For unstructured data, hand-crafted feature engineering is crucial
• Development of constrained models to control complexity and ensure interpretability
• E.g., scoring models, GAMs, subscale models
• Direct verification of the model due to constrained model structures
• No need for additional explanations via approximations or simplifications
• Adjustments directly in the model as a single source of truth, affecting the predicted outcome and model insights
• Assessment limited to the evaluation of the model quality
• No need for additional evaluation of explanation fidelity
Inherently Interpretable Machine Learning


Traditional
P
Q
ω1 ω2 … ωn
φ1 φ2 … φn
[R,t]
Target Selection
Feature Extraction
Feature Matching
Pose Optimization
Learning-based
Inputs
Point Cloud
2D Image
Voxel Grid
Feature Vector
Mesh
Matching Vector
Conceptual Space
φ1 τ1 ν1
φ2 τ2 ν2
φn τn νn
Outputs
Point Cloud
2D Image
Feature Vector
Voxel Grid
Matching Vector
Mesh
Transformation
[R,t]


Cognitive Dimension
Technical Dimension
Explainability
From a cognitive perspective, explainability refers to the ability to accurately describe the mathematical decision logic, technical process, mechanism, or implementation that led to a model's output. The focus is on the property of the model itself.
Supporting references: Berente et al. (2021), Broniatowski (2021)
From a technical perspective, (post-hoc) explainability implies that a model is so complex that it requires auxiliary explanation tools or methods to make its decision logic comprehensible for humans.
Supporting references: Burkart & Huber (2021), Meske et al. (2022), Rudin (2019)
General definition:
In the context of AI and ML models, interpretability and explainability generally refer to a model's ability to explain or present its behavior in human-understandable terms.
Supporting references: Doshi-Velez & Kim (2017), Guidotti et al. (2018)
Interpretability
From a cognitive perspective, interpretability refers to the ability to derive meaning from a model's output so that humans can make a decision by leveraging their background knowledge for a specific case. The focus is on the human interpreter's role.
Supporting references: Berente et al. (2021), Broniatowski (2021)
From a technical perspective, (inherent) interpretability denotes that a model is already structured in a way that its decision logic is inherently accessible and understandable for humans without the need for post-hoc explanation.
Supporting references: Burkart & Huber (2021), Meske et al. (2022), Rudin (2019)