SOI Matters: Analyzing Multi-Setting Training Dynamics in Pretrained Language Models via Subsets of Interest

Shayan Vassef[1]*, **Amirhossein Dabiriaghdam**[2]*, Mohammadreza Bakhtiari[3]*, Yadollah Yaghoobzadeh[4]
[1]University of Illinois Chicago, [2]University of British Columbia, [3]Stony Brook University, [4]University of Tehran
Equal Contribution

Introduction

- Despite LLMs' prominence, smaller fine-tuned PLMs (e.g., BERT, XLM-RoBERTa) often match or exceed LLM performance on specialized NLP tasks
- Multi-setting learning approaches (multi-task, multi-source, multi-lingual) offer enhanced generalization but their training dynamics remain poorly understood
- **Key Challenge**: Understanding how training examples behave differently under single vs multi-setting configurations and leveraging these insights for improved out-of-distribution (OOD) performance

Subsets of Interest (SOI) Framework

We introduce SOI, a novel categorization framework that identifies **6 distinct learning behavior patterns** based on per-example trajectories over 10 training epochs:

- **UNE (Unlearned Examples)**: Never learned after initial attempts - samples the model consistently fails to classify correctly
- **ACE (Always Correct Examples)**: Consistently correct throughout all epochs - easy-to-learn samples
- **1t-FRGE (1-time Forgettable)**: Exactly one forgetting event followed by recollection
- **$\geq$2t-FRGE (Multiple Forgettable)**: Multiple forgetting and recollection cycles - highly unstable learning
- **ELE (Early-Learned Examples)**: Learned and stabilized by epoch 5 - moderately easy samples
- **LLE (Late-Learned Examples)**: Learned and stabilized after epoch 5 - challenging but eventually learnable samples

Key Insight: By tracking how examples transition between SOI categories when moving from single-setting to multi-setting training, we gain actionable insights for optimizing model performance.

Experimental Setup: Three Parallel Comparisons

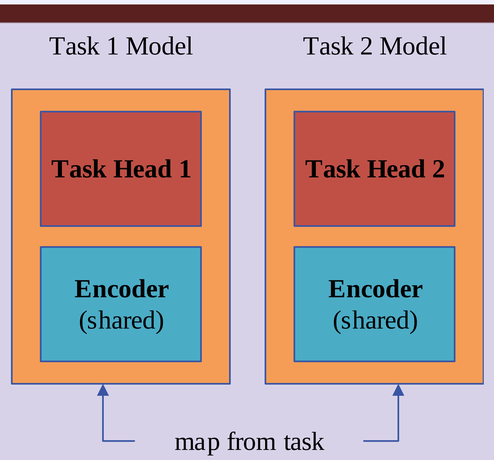
1. Multi-task vs Single-task Learning (English):
- **Entailment**: SciTail (ID: 23K) $\rightarrow$ RTE (OOD: 2.5K)
- **Paraphrase**: MRPC (ID: 3.7K) $\rightarrow$ QQP (OOD: 4K)
- **Sentiment**: Twitter Airline (ID: 8K) $\rightarrow$ SST-2 (OOD: 4K)
- Task pairs: SE, SP, PE (similar vs dissimilar tasks)

2. Multi-source vs Single-source Learning (Sentiment):
- **IMDB**: Movie reviews (ID: 40K training)
- **Yelp**: Business reviews (ID: 40K training)
- **Sentiment140**: Twitter content (ID: 40K training)
- All sources $\rightarrow$ SST-2 (OOD: 5K)
- Source pairs: IY, SY, IS

3. Multi-lingual vs Single-lingual Learning (Intent):
- **English**: CLINC150-Small (ID: 7.6K)
- **French**: MIAM-LORIA (ID: 8.5K)
- **Persian**: MASSIVE subset (ID: 11.5K)
- All languages $\rightarrow$ Burmese translations (OOD)
- Language pairs: En-Fr, En-Fa, Fr-Fa

Architecture: Shared encoder (BERT for tasks/sources, XLM-R for languages) with setting-specific classification heads

Unified Architecture



Multi-setting architecture with shared encoder and task/source/language-specific heads.

From Architecture to Visualization: The SOI Analysis Pipeline

Step 1: Training & Tracking
- Train models for 10 epochs recording per-example correctness each epoch
- Track binary learning status: correct (1) vs incorrect (0) predictions

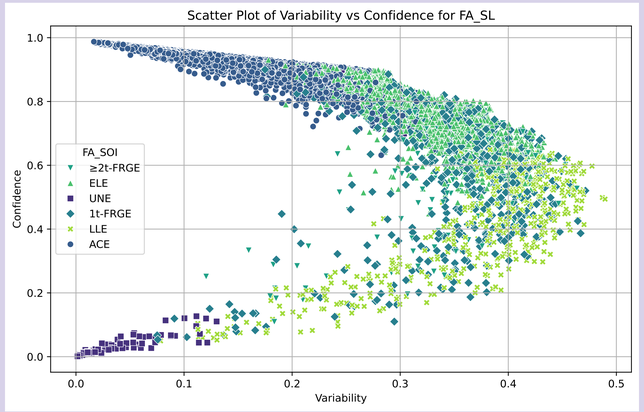
Step 2: SOI Classification
- Analyze learning trajectories to assign each example to one of 6 SOI categories
- Example: [0,1,0,1,0,1,0,0,0,0] $\rightarrow$ $\geq$2t-FRGE (multiple forget/recollect cycles)

Step 3: Cartography Visualization
- Map examples in confidence-variability space
- **Confidence**: Average of highest prediction probabilities across epochs
- **Variability**: Standard deviation of predictions across epochs
- Reveals easy-to-learn, hard-to-learn, and ambiguous regions

Step 4: Transition Analysis
- Track how examples migrate between SOI categories from single $\rightarrow$ multi-setting
- Use heatmaps to visualize transition patterns
- Identify stable vs unstable learning behaviors

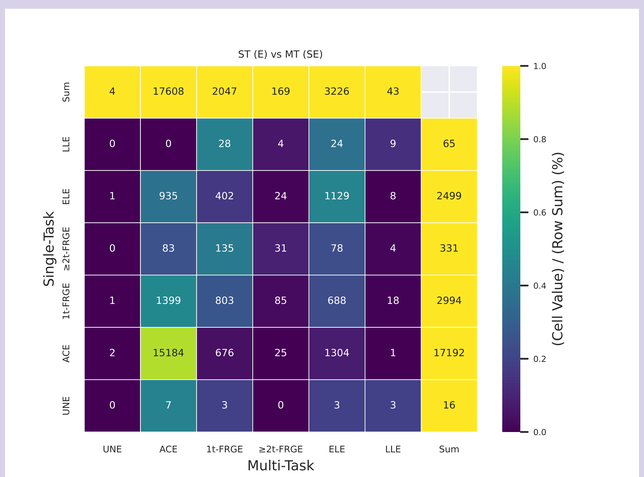
Step 5: Strategic Fine-tuning
- Select subsets based on SOI transitions for second-stage fine-tuning
- Target examples showing positive transfer potential

Dataset Cartography: SOI Distribution



Confidence vs variability map for Persian single-lingual learning. UNE concentrates in hard-to-learn region (low confidence, low variability), ACE in easy-to-learn region (high confidence, low variability), while forgettable examples (1t-FRGE, $\geq$2t-FRGE) spread across ambiguous regions with high variability.

SOI Transition Heatmaps



Example migration between SOI categories from single-task (E) to multi-task (SE). Each cell shows the count of examples transitioning from one SOI category to another, revealing how multi-setting training affects learning dynamics.

Comprehensive Results: First & Second Stage Fine-tuning

Multi-Source Learning (Sentiment Analysis):

Model & Source	ID	OOD-1st	OOD-2nd
Single: IMDB	89.4	79.4	–
Single: Yelp	93.8	79.6	–
Single: Sentiment140	82.7	76.0	–
Multi (IY): IMDB	90.2	**83.9** (+4.5)	83.6
Multi (IY): Yelp	94.1	**84.3** (+4.7)	84.1
Multi (SY): Sentiment140	83.6	**79.0** (+3.0)	**79.4**
Multi (SY): Yelp	93.7	**83.2** (+3.6)	82.7
Multi (IS): IMDB	90.2	**85.5** (+6.1)	84.9
Multi (IS): Sentiment140	83.5	**83.0** (+7.0)	83.1

Multi-Task Learning (English):

Model & Task	ID	OOD-1st	OOD-2nd
Single: Entailment	89.3	43.9	–
Single: Sentiment	94.6	76.7	–
Single: Paraphrase	81.7	62.7	–
Multi (SP): Sentiment	95.0	75.3 (-1.4)	74.4
Multi (SP): Paraphrase	80.3	57.3 (-5.4)	**58.8**
Multi (SE): Sentiment	95.1	62.7 (-14.0)	**64.9**
Multi (SE): Entailment	91.9	38.6 (-5.3)	38.2
Multi (PE): Paraphrase	79.3	**69.6** (+6.9)	**70.0**
Multi (PE): Entailment	89.6	**45.7** (+1.8)	45.1

Multi-Lingual Learning (Intent Classification):

Model & Language	ID	OOD-1st	OOD-2nd
Single: English	84.5	52.8	–
Single: French	88.5	49.0	–
Single: Persian	87.4	62.9	–
Multi (En-Fr): English	84.4	51.9 (-0.9)	51.8
Multi (En-Fr): French	88.7	41.6 (-7.4)	40.9
Multi (En-Fa): English	84.7	48.0 (-4.8)	48.1
Multi (En-Fa): Persian	87.4	**63.3** (+0.4)	**63.6**
Multi (Fr-Fa): French	89.4	**52.2** (+3.2)	52.2
Multi (Fr-Fa): Persian	87.2	61.0 (-1.9)	**61.4**

Key Insights & Why SOI Matters

First-Stage Findings:
- **Multi-source**: Consistently improves OOD (up to +7%) when task and language are fixed - diverse data sources enhance robustness
- **Multi-task**: Benefits depend on task similarity - related tasks (Paraphrase-Entailment) show mutual gains (+6.9%, +1.8%), while dissimilar tasks can hurt performance
- **Multi-lingual**: Mixed, language-dependent effects - benefits are asymmetric and not reciprocal

Why SOI-Guided Fine-tuning Helps:
- **Targeted stability**: Prioritizes examples showing positive transfer across settings based on transition patterns
- **Noise-aware selection**: Avoids overfitting on ACE (too easy) or high-churn examples (too unstable)
- **Transfer-based strategy**: Uses empirical evidence of cross-setting behavior rather than static difficulty metrics

Second-Stage Results:
- Most effective in multi-task setting, recovering performance where first-stage dropped
- Limited gains in multi-source (already near-optimal)
- Stable improvements in select multi-lingual configurations

Conclusions & Future Directions

Main Contributions:
- Introduced SOI framework with 6 fine-grained learning behavior categories
- Demonstrated that multi-setting benefits are highly conditional on task/source/language relationships
- Showed SOI-guided two-stage fine-tuning provides targeted OOD improvements

Future Work:
- Extend SOI analysis to decoder-based LLMs (GPT-style models)
- Explore beyond pairwise combinations (3+ simultaneous settings)
- Investigate curriculum learning strategies based on SOI categories

Paper (ArXiv)

Code (GitHub)