Supplementary Material for Paper "Do LLMs Learn Structure or Names? A Study on the Robustness of Design-Pattern Detection to Identifier Shifts"

Ichsan Budiman
ixb402@student.bham.ac.uk
School of Computer Science, University of Birmingham
Birmingham, United Kingdom

Leandro L. Minku
l.l.minku@bham.ac.uk
School of Computer Science, University of Birmingham
Birmingham, United Kingdom

ACM Reference Format:
Ichsan Budiman and Leandro L. Minku. 2026. Supplementary Material for Paper "Do LLMs Learn Structure or Names? A Study on the Robustness of Design-Pattern Detection to Identifier Shifts". In *22nd International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE '26), July 05, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3803846.3807465

This supplementary material provides additional experimental details, hyperparameter configurations, dataset statistics, and full cross-dataset evaluation results referenced in the main paper.

1 Experimental Setup Details

1.1 Dataset Distribution

Table 1 lists the number of instances per design-pattern class used in our experiments.

Table 1: Distribution of design-pattern instances across the 13 pattern classes.

Pattern	Count	Pattern	Count
Adapter	100	Visitor	100
Facade	100	FactoryMethod	100
Observer	100	Proxy	100
None	100	Memento	100
Builder	100	Prototype	100
Decorator	100	Singleton	100
AbstractFactory	100		

1.2 Fine-tuning and Embedding Extraction

Table 2 summarises the hyperparameters shared across all pretrained language model (PLM) fine-tuning runs.

Table 2: Hyperparameter configuration used for fine-tuning all PLMs.

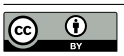
Parameter	Value
Epochs	10
Batch size	16
Learning rate	5×10^{-5}
Optimizer	AdamW
Scheduler	linear decay
Warmup steps	0
Weight decay	0.01
Dropout	model default
Max sequence length	512

1.3 Downstream Classifier Hyperparameters

Table 3 lists the fixed hyperparameters for each downstream classifier paired with PLM embeddings.

Table 3: Downstream ML classifiers and fixed hyperparameters used with PLM embeddings.

Classifier	Hyperparameters
SVM (RBF)	`C=1000`; `kernel=rbf`; `gamma=scale`; `class_weight=balanced`; `max_iter=100000`; `probability=True`
Random Forest	`n_estimators=100`; `class_weight=balanced`
AdaBoost (SAMME)	`base estimator: RF(n_estimators=100)`; `learning_rate=1`; `algorithm=SAMME`
MLP	`hidden_layer_sizes=(100,)`; `activation=relu`; `solver=adam`; `learning_rate=adaptive`; `max_iter=500`; `alpha=`10^{-4}; `early_stopping=True`; `validation_fraction=0.1`; `batch_size=auto`



PROMISE '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2584-5/2026/07
https://doi.org/10.1145/3803846.3807465

1.4 Kernel SHAP Configuration

Table 4 details the Kernel SHAP settings used throughout the explainability analysis.

Table 4: Kernel SHAP configuration used for explainability analysis.

Component	Value
GPU type	NVIDIA A100
Background k-means k	30
Background max samples	1300 (per variant)
Background seed	[42, 123, 456, 789, 2025]
SHAP `n_samples`	30
SHAP L_1 regulariser	`num_features(30)`
L_1 reg. multiplier	2.5
Max sequence length	512
Batch size (SHAP)	1
Samples per pattern	100% of instances per class
Tokenizer type	Native (same as fine-tuning)

1.5 Two-Bucket SHAP: Syntax vs. Identifiers

Tables 5–7 report token-level SHAP bucket statistics (fraction of tokens where syntax wins, identifier wins, or ties) per design-pattern class for the RAW, CLEAN, and ANON variants, respectively.

Table 5: Token-level SHAP bucket statistics per design-pattern class for the RAW variant.

Pattern	Syntax Wins	Identifier Wins	Tie
AbstractFactory	0.0834	0.9096	0.0069
Adapter	0.0570	0.9427	0.0003
Builder	0.0767	0.9233	0.0000
Decorator	0.0958	0.8984	0.0058
Facade	0.0808	0.9172	0.0020
FactoryMethod	0.0687	0.9295	0.0018
Memento	0.1234	0.8766	0.0000
None	0.1439	0.8492	0.0069
Observer	0.0824	0.9044	0.0131
Prototype	0.0574	0.9415	0.0011
Proxy	0.0950	0.9022	0.0028
Singleton	0.0736	0.9219	0.0045
Visitor	0.0697	0.9285	0.0018
OVERALL	0.0815	0.9152	0.0033

Table 6: Token-level SHAP bucket statistics per design-pattern class for the CLEAN variant.

Pattern	Syntax Wins	Identifier Wins	Tie
AbstractFactory	0.1354	0.8469	0.0177
Adapter	0.1708	0.8187	0.0105
Builder	0.1665	0.8275	0.0059
Decorator	0.1820	0.8065	0.0115
Facade	0.1788	0.8120	0.0092
FactoryMethod	0.1325	0.8550	0.0126
Memento	0.1562	0.8313	0.0125
None	0.1945	0.7821	0.0233
Observer	0.1347	0.8152	0.0501
Prototype	0.1465	0.8446	0.0088
Proxy	0.2110	0.7783	0.0107
Singleton	0.1613	0.8073	0.0314
Visitor	0.1405	0.8388	0.0207
OVERALL	0.1635	0.8206	0.0159

Table 7: Token-level SHAP bucket statistics per design-pattern class for the ANON variant.

Pattern	Syntax Wins	Identifier Wins	Tie
AbstractFactory	0.4789	0.5134	0.0076
Adapter	0.5202	0.4741	0.0058
Builder	0.5516	0.4460	0.0024
Decorator	0.4979	0.4979	0.0041
Facade	0.5052	0.4918	0.0031
FactoryMethod	0.5308	0.4537	0.0155
Memento	0.5444	0.4495	0.0061
None	0.6509	0.3432	0.0059
Observer	0.4780	0.4962	0.0258
Prototype	0.5688	0.4251	0.0061
Proxy	0.5105	0.4834	0.0060
Singleton	0.5041	0.4878	0.0081
Visitor	0.5142	0.4757	0.0101
OVERALL	0.5133	0.4782	0.0086

2 Additional Results

2.1 Baseline Performance: Trained and Evaluated on RAW

Figure 1 reports the best PLM-classifier combination when both training and evaluation use the RAW variant.

2.2 Feature Reliance by Training Variant

Figure 2 shows aggregated syntax and identifier reliance ratios per PLM, grouped by training variant. Figure 3a and 3b shows per-pattern recall heatmaps under RAW→RAW and RAW→ANON evaluation conditions.

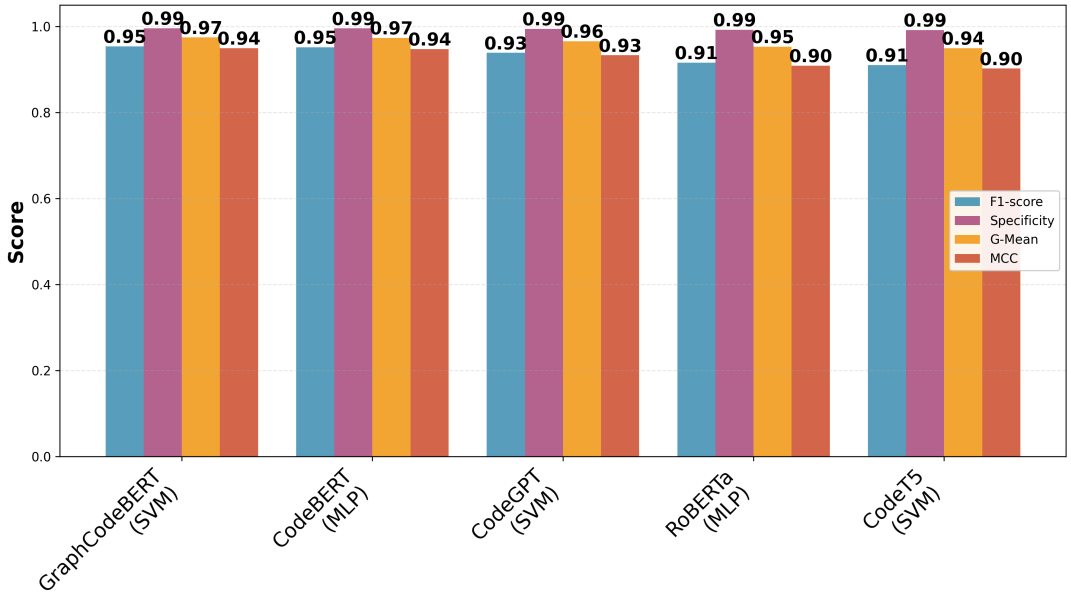


Figure 1: Best PLM-classifier combination performance when trained and evaluated on RAW data.

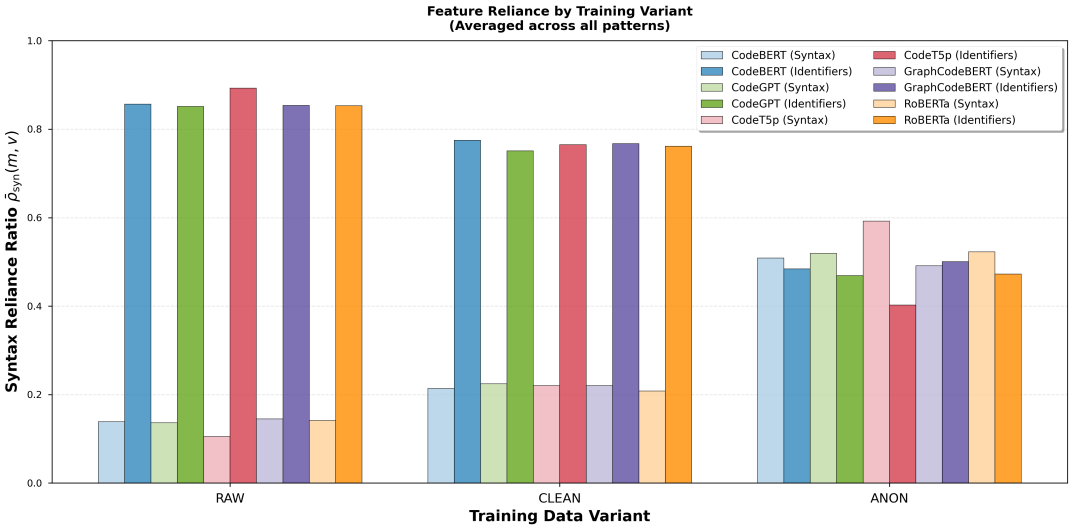
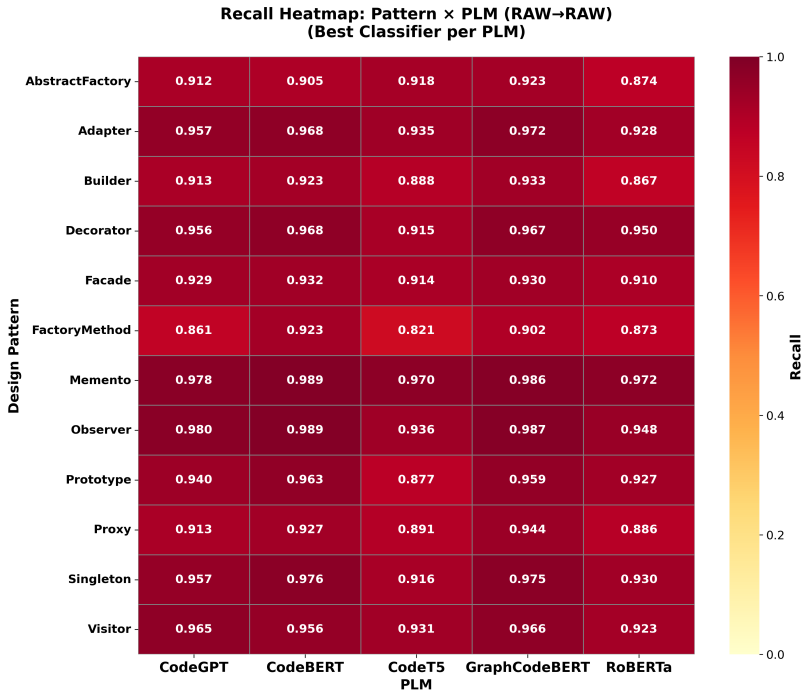
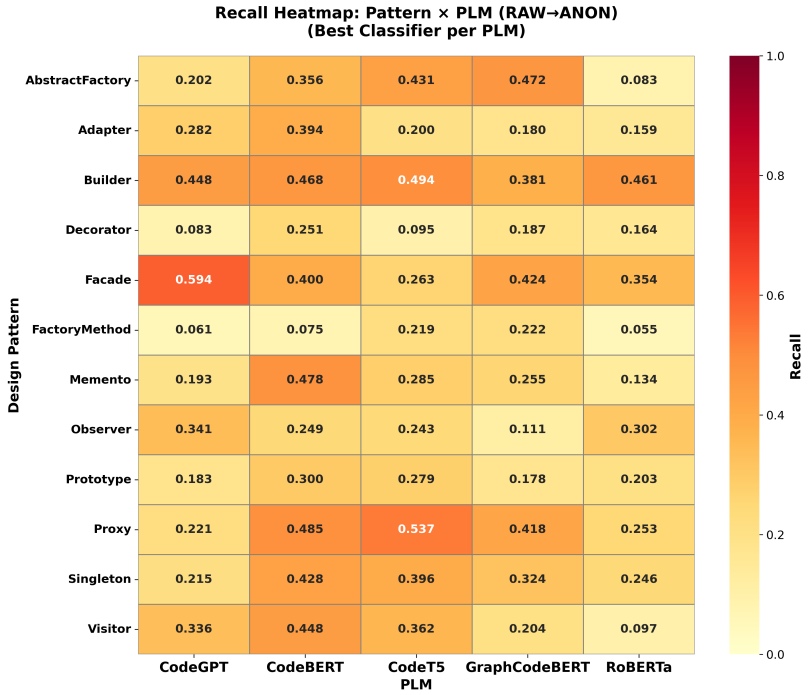


Figure 2: Aggregated syntax ($\bar{\rho}_{syn}$) and identifier ($\bar{\rho}_{id}$) reliance ratios per PLM, grouped by training variant.

2.3 PLMs Performance under RAW and ANON Evaluation



(a) RAW→RAW.



(b) RAW→ANON.

(c) Per-pattern recall heatmaps (best classifier per PLM) under RAW and ANON evaluation.

2.4 Full Cross-Dataset Evaluation Results

Tables 8 and 9 report the complete cross-dataset evaluation results across all PLM–classifier combinations and training/evaluation variant pairs. Table 8 presents macro-F1 and MCC scores, while Table 9 reports G-Mean and Specificity. Bold entries indicate the highest score within each PLM group. Results are shown for all nine train/eval combinations spanning the RAW, CLEAN, and ANON variants.

Table 8: Cross-dataset results. Top half reports macro-F1, bottom half reports MCC. Bold indicates the highest score within each PLM group.

PLM	Cls.	Train RAW			Train CLEAN			Train ANON		
		Eval RAW	Eval CLEAN	Eval ANON	Eval RAW	Eval CLEAN	Eval ANON	Eval RAW	Eval CLEAN	Eval ANON
					Macro-F1					
CodeGPT	ada	0.931	0.927	0.274	0.860	0.955	0.267	0.443	0.647	0.715
	mlp	0.928	0.928	0.257	0.856	0.945	0.233	0.445	0.661	0.722
	rf	0.931	0.926	**0.279**	**0.878**	0.956	**0.296**	0.440	0.652	0.715
	svm	**0.939**	**0.934**	0.269	0.847	**0.959**	0.295	**0.513**	**0.724**	**0.773**
codebert-base	ada	0.951	**0.942**	0.316	0.912	0.962	0.352	0.595	0.705	0.763
	mlp	**0.952**	0.941	**0.358**	0.908	0.961	**0.367**	0.591	0.705	0.754
	rf	0.951	0.942	0.325	**0.913**	0.961	0.353	0.597	0.708	0.765
	svm	0.951	0.942	0.314	0.913	**0.962**	0.362	**0.606**	**0.711**	**0.771**
codet5p	ada	0.847	0.685	0.270	0.672	0.777	0.320	0.532	0.587	0.665
	mlp	0.882	**0.811**	0.311	0.766	0.840	**0.347**	**0.626**	**0.677**	0.690
	rf	0.848	0.689	0.272	0.673	0.783	0.322	0.533	0.589	0.662
	svm	**0.910**	0.805	**0.314**	**0.778**	**0.866**	0.338	0.614	0.663	**0.727**
graphcodebert-base	ada	0.949	0.940	0.261	0.894	0.961	0.235	0.650	0.755	0.779
	mlp	0.946	0.942	**0.316**	0.901	0.959	**0.278**	0.661	0.752	0.772
	rf	0.950	0.942	0.270	0.898	0.961	0.238	0.652	0.752	0.781
	svm	**0.954**	**0.946**	0.279	**0.905**	**0.966**	0.243	**0.678**	**0.766**	**0.787**
roberta-base	ada	0.895	0.840	0.188	0.697	0.800	0.203	0.299	0.412	0.517
	mlp	**0.916**	**0.881**	**0.206**	**0.750**	**0.813**	**0.214**	**0.315**	**0.427**	0.500
	rf	0.895	0.843	0.192	0.697	0.797	0.204	0.295	0.412	**0.519**
	svm	0.866	0.824	0.179	0.690	0.767	0.195	0.288	0.390	0.488
					MCC					
CodeGPT	ada	0.925	0.920	0.229	0.847	0.950	0.224	0.392	0.614	0.688
	mlp	0.921	0.921	0.200	0.841	0.939	0.181	0.398	0.628	0.696
	rf	0.924	0.919	**0.234**	**0.866**	0.952	**0.255**	0.390	0.620	0.688
	svm	**0.933**	**0.927**	0.215	0.828	**0.955**	0.254	**0.472**	**0.695**	**0.751**
codebert-base	ada	0.946	**0.937**	0.259	0.906	0.957	0.302	0.550	0.672	0.739
	mlp	**0.947**	0.936	**0.309**	0.903	0.956	**0.317**	0.551	0.675	0.730
	rf	0.947	0.937	0.269	**0.907**	0.957	0.302	0.552	0.676	0.741
	svm	0.946	0.937	0.260	0.907	**0.958**	0.312	**0.562**	**0.679**	**0.747**
codet5p	ada	0.832	0.665	0.210	0.645	0.756	0.259	0.496	0.552	0.634
	mlp	0.871	**0.796**	0.253	0.748	0.825	**0.293**	**0.598**	**0.648**	0.662
	rf	0.833	0.670	0.211	0.648	0.761	0.261	0.499	0.555	0.631
	svm	**0.901**	0.787	**0.259**	**0.761**	**0.852**	0.289	0.579	0.629	**0.702**
graphcodebert-base	ada	0.944	0.935	0.202	0.886	0.957	0.198	0.617	0.729	0.757
	mlp	0.941	0.937	**0.260**	0.896	0.955	**0.234**	0.630	0.727	0.751
	rf	0.945	0.937	0.209	0.891	0.957	0.198	0.621	0.725	0.759
	svm	**0.949**	**0.941**	0.221	**0.898**	**0.962**	0.204	**0.649**	**0.741**	**0.767**
roberta-base	ada	0.885	0.827	0.124	0.671	0.782	0.129	0.241	0.356	0.471
	mlp	**0.908**	**0.872**	**0.144**	**0.731**	**0.792**	**0.138**	**0.259**	**0.374**	0.453
	rf	0.885	0.829	0.127	0.671	0.778	0.131	0.236	0.356	**0.474**
	svm	0.852	0.808	0.113	0.660	0.744	0.126	0.228	0.335	0.443

Table 9: Cross-dataset results. Top half reports G-Mean, bottom half reports Specificity. Bold indicates the highest score within each PLM group.

PLM	Cls.	Train RAW			Train CLEAN			Train ANON		
		Eval RAW	Eval CLEAN	Eval ANON	Eval RAW	Eval CLEAN	Eval ANON	Eval RAW	Eval CLEAN	Eval ANON
G-Mean										
CodeGPT	ADA	0.962	0.959	0.492	0.918	0.975	0.487	0.630	0.783	0.831
	mlp	0.960	0.959	0.472	0.916	0.969	0.448	0.632	0.794	0.837
	RF	0.962	0.958	**0.496**	**0.928**	0.976	**0.512**	0.629	0.786	0.831
	SVM	**0.966**	**0.963**	0.470	0.905	**0.977**	0.503	**0.688**	**0.835**	**0.868**
CodeBERT-base	ADA	0.973	**0.968**	0.514	0.949	0.979	0.559	0.746	0.822	0.862
	mlp	**0.973**	0.967	**0.570**	0.947	0.978	**0.575**	0.745	0.823	0.857
	RF	0.973	0.967	0.521	**0.949**	0.979	0.559	0.746	0.824	0.863
	SVM	0.973	0.968	0.514	0.949	**0.979**	0.565	**0.754**	**0.826**	**0.867**
CodeT5	ADA	0.912	0.813	0.495	0.801	0.871	0.537	0.696	0.739	0.800
	mlp	0.933	**0.892**	0.519	0.862	0.909	**0.560**	**0.765**	**0.803**	0.817
	RF	0.913	0.816	0.497	0.803	0.874	0.540	0.697	0.741	0.798
	SVM	**0.950**	0.886	**0.531**	**0.869**	**0.924**	0.553	0.758	0.794	**0.841**
GraphCodeBERT-base	ADA	0.972	0.967	0.484	0.938	0.978	0.445	0.786	0.854	0.872
	mlp	0.970	0.968	**0.533**	0.943	0.977	**0.496**	0.793	0.853	0.868
	RF	0.972	0.967	0.492	0.940	0.979	0.446	0.787	0.852	0.873
	SVM	**0.974**	**0.970**	0.497	**0.944**	**0.981**	0.450	**0.805**	**0.862**	**0.877**
RoBERTa-base	ADA	0.941	0.909	0.401	0.820	0.886	0.420	0.523	0.620	0.700
	mlp	**0.953**	**0.932**	**0.419**	**0.854**	**0.891**	**0.433**	**0.538**	**0.631**	0.687
	RF	0.941	0.910	0.404	0.820	0.884	0.422	0.520	0.620	**0.702**
	SVM	0.924	0.900	0.391	0.816	0.865	0.415	0.515	0.605	0.681
Specificity										
CodeGPT	ADA	0.994	0.993	0.935	0.987	0.996	0.935	0.949	0.968	0.974
	mlp	0.993	0.993	0.933	0.987	0.995	0.931	0.949	0.969	0.975
	RF	0.994	0.993	**0.936**	**0.989**	0.996	**0.938**	0.949	0.968	0.974
	SVM	**0.994**	**0.994**	0.933	0.985	**0.996**	0.937	**0.955**	**0.974**	**0.979**
CodeBERT-base	ADA	0.996	**0.995**	0.937	0.992	0.996	0.941	0.962	0.972	0.978
	mlp	**0.996**	0.995	**0.942**	0.992	0.996	**0.943**	0.962	0.973	0.978
	RF	0.996	0.995	0.938	**0.992**	0.996	0.941	0.962	0.973	0.978
	SVM	0.996	0.995	0.937	0.992	**0.997**	0.942	**0.963**	**0.973**	**0.979**
CodeT5	ADA	0.986	0.972	0.934	0.970	0.980	0.938	0.957	0.962	0.969
	mlp	0.989	**0.983**	0.937	0.979	0.985	**0.941**	**0.966**	**0.970**	0.972
	RF	0.986	0.972	0.934	0.970	0.980	0.938	0.958	0.962	0.969
	SVM	**0.992**	0.982	**0.938**	**0.980**	**0.988**	0.940	0.964	0.969	**0.975**
GraphCodeBERT-base	ADA	0.995	0.995	0.933	0.990	0.996	0.932	0.968	0.977	0.980
	mlp	0.995	0.995	**0.938**	0.991	0.996	**0.936**	0.969	0.977	0.979
	RF	0.995	0.995	0.934	0.991	0.996	0.932	0.968	0.977	0.980
	SVM	**0.996**	**0.995**	0.935	**0.991**	**0.997**	0.932	**0.971**	**0.978**	**0.981**
RoBERTa-base	ADA	0.990	0.986	0.927	0.973	0.982	0.927	0.937	0.946	0.956
	mlp	**0.992**	**0.989**	**0.928**	**0.977**	**0.983**	**0.928**	**0.938**	**0.948**	0.954
	RF	0.990	0.986	0.927	0.973	0.982	0.927	0.936	0.946	**0.956**
	SVM	0.988	0.984	0.926	0.972	0.979	0.927	0.936	0.945	0.954