Paper	NL2Type: Inferring JavaScript function types from natural language information
ID	AP5
Experiments	4
Comments	2 non-comparative experiments (RQ1 and RQ5) and a qualitative (RQ3)

Aspect	Element	E1	Comments	Code
Experiment type		Optimization+Evaluation	Not sure what it is	
Hypotheses	Research hypotheses	Yes	RQ1, RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 3 #neurons/layer: l1 (100), l2 (256), l3 (1000) connections: input, bi-directional LSTM, fully connected, output (softmax) activation functions: No params. Initialization: No	Embedding built upon Word2Vec, apparently being re-trained, but not clear if it is re-trained separately or together with DNN proposed. I have excluded from the paper the model hyperparameters of Word2Vec, as its architecture is not described (word embedding size:100, context size:5, min. ocurrence of word: 5) #neurons deduced	**File "model.h5" provided. When loaded in python:** #layers: 3 connections: input, bidirectional LSTM, dense #neurons/layer: 100,**512,978** **Activation functions and params. Initialization unknown**
	Model parameters	biases: No weights: No	But they are in the artefact.	**Yes**
	DL algorithm	representation: Yes model type: Yes loss function: categorical cross entropy regularization: dropout (20%) optimization: Adam (defaults?)		**Code not provided**
	Training hyperparameters	train-test split: 80-20 learning rate: No #iterations: No batch size: 256 #epochs: 12	No need of K-cross validation, due to large amount of data	**Code not provided**
	Training data	Yes	Linked to artifact	**Provided (dropbox link)**
Operationalization	Factors and treatments	DNN Model	NL2Type, NL2Type w/o comments, naive (always same answer, k most common types)	
	Response variable, elaboration and metric	Yes	Precision, recall, F1 on top-1-3-5 predicted Efficiency (average time perfunction or total) for NL2Type	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Linked to artifact	**Test set provided (dropbox link)**
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory.		
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	No		

Paper	NL2Type: Inferring JavaScript function
ID	AP5
Experiments	4
Comments	2 non-comparative experiments (RQ1

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Seems uses the "optimized" version	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as next?		Same as previous
	Model parameters	Same as next?		Same as previous
	DL algorithm	Same as next?		Same as previous
	Training hyperparameters	Same as next?		Same as previous
	Training data	Yes	Linked to artifact	Same as previous
Operationalization	Factors and treatments	DNN Model	DeepTyper (JSNice is not a DNN), NL2Type For DeepTyper, use their publicly available artifact, and do not apply confidence threshold	
	Response variable, elaboration and metric	Yes	Precision, recall, F1 on top-1 predicted	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	Partially	Create a front-end forNL2Type to use dataset in DeepTyper and allow fair comparison	
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Linked to artifact	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory.		
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	No		

Paper	NL2Type: Inferring JavaScript function
ID	AP5
Experiments	4
Comments	2 non-comparative experiments (RQ1

Aspect	Element	E3	Comments	Code
Experiment type		Evaluation	Evaluate DNN for another task (inconsistencies detection)	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as next?		Same as previous
	Model parameters	Same as next?		Same as previous
	DL algorithm	Same as next?		Same as previous
	Training hyperparameters	Same as next?		Same as previous
	Training data	Yes	Linked to artifact	Same as previous
Operationalization	Factors and treatments	DNN	1 treatment only	
	Response variable, elaboration and metric	Yes	Frequency of potential inconsistency types (inconsistency/non-standard type annotation/misclassification)	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	Partially		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	Partially	Multiple runs to check the predictions. Some neurons are purposefully deactivated during prediction.	
	Number of experimental units	No	Seems more than 1	
Instrumentation	Test set	Yes	Linked to artifact	Same as previous
	Measuring instruments	Yes		
	Measurement procedure	Yes	NL2Type is used in a different way. The return value is used to check if the predicted type matches the real one.	
	Technological infrastructure	Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory.		
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	No		

Paper	NL2Type: Inferring JavaScript function
ID	AP5
Experiments	4
Comments	2 non-comparative experiments (RQ1

Aspect	Element	E4	Comments	Code
Experiment type		Optimization	It should be the first experiment...	
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 4 #neurons/layer: l1 (?), l2 (100?), l3 (256?), l4 (1000?) connections: input, bi-directional LSTM, fully connected, output (softmax) activation functions: No params. Initialization: No	Embedding built upon Word2Vec, apparently being re-trained, but not clear if it is re-trained separately or together with DNN proposed. I have excluded from the paper the model hyperparameters of Word2Vec, as its architecture is not described (word embedding size:100, context size:5, min. ocurrence of word: 5) #neurons deduced	Same as previous
	Model parameters	biases: No weights: No	But they are in the artefact.	Same as previous
	DL algorithm	representation: Yes model type: Yes loss function: categorical cross entropy regularization: dropout (20%) optimization: Adam (defaults?)		Same as previous
	Training hyperparameters	train-test split: 80-20 learning rate: No #iterations: No batch size: 256 #epochs: 12	No need of K-cross validation, due to large amount of data	Same as previous
	Training data	Yes	Linked to artifact	Same as previous
Operationalization	Factors and treatments	DNN architecture	output of DNN (5...5000) Paper mentions they have run experiments to choose hyperparameters, but they are not described	
	Response variable, elaboration and metric	Yes	Precision, recall, F1 on top-1 predicted	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	Partially	Input representation: words in names: 6 words in comment: 12 words in comment: 10 #pars: 10	
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No		
Instrumentation	Test set	Yes	Linked to artifact	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Implemented in Python Preprocessing: Python NLTK library Word2Vec DNN: Keras Ubuntu 16.04 computer with Intel Xeon E5-2650 with 48 cores, 64GB memory and NVIDIA Tesla P100 GPU with 16GB of memory.		
Population	Objects (chars. of the experimental datasets)	No	Mentions JavaScript files/libraries	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	No		

Paper A novel neural source code representation based on abstract syntax tree
ID AP8
Experiments 4
Comments

Aspect	Element	E1	Comments	Code
Experiment type		Evaluation	Source code classification	
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 6? #neurons/layer: GRU (100) connections: encoder, recurrent, pooling, output activation functions: Some mentioned: identity (encoder) params. Initialization: No	For encodder mentions Word2Vec. Its role not explained. Could be the pre-trained encoder (mesning weights are initialized with these values)	#neurons/layer: GRU (100) Encode dimension: 128 Activation function: relu
	Model parameters	biases: Yes weights: Yes	Explicitly says that "trained models are stored"	Cannot see them
	DL algorithm	representation: Yes model type: Yes loss function: cross-entropy regularization: No optimization: AdaMax		loss function: cross-entropy but also BCELoss regularization: dropout (0.2) optimization: AdaMax
	Training hyperparameters	train-test split: 60-20-20 learning rate: 0.002 #iterations: No batch size: 64 #epochs: max. 15		train-test split: not clear. Not understandable formula learning rate: Not found #iterations: No batch size: 64 #epochs: max. 15
	Training data	Yes	OJ. Referenced	
Operationalization	Factors and treatments	Partially	ASTNN, TextCNN, LSTM, LSCNN For other approaches: TextCNN: kernel size=3, filters=100 LSTM: hidden states =100 LSCNN: nothing	
	Response variable, elaboration and metric	Yes	Accuracy	
Design	Design type	No	Seems 1 factor-6 treatment (TextCNN, LSTM, TBCNN,LSCNN,PGD+GGNN)	
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure			
	Number of experimental units	No	Assume 1	
Instrumentation	Test set	Yes	OJ. Referenced	
	Measuring instruments	Yes		
	Measurement procedure	Yes		
	Technological infrastructure	Partially	pycparser (C) and javalang (Java) to obtain ASTs train embeddings using word2vec (embedding size=128) 16 cores of 2.4GHz CPU, Titan Xp GPU	
Population	Objects (chars. of the experimental dataset)	Partially	Mention the datasets and references (OJ)	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	A novel neural source code represent
ID	AP8
Experiments	4
Comments	

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Code clone detection	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		**Same as previous**
	Model parameters	Same as previous?		**Same as previous**
	DL algorithm	Same as previous?	Loss function is binary cross-entropy	**Same as previous**
	Training hyperparameters	Same as previous?	#epochs: max. 5 Threshold: 0.5	**Same as previous**
	Training data	Yes	OJ, BCB, referenced	
Operationalization	Factors and treatments	Partially	ASTNN, RAE+, CDLH For other approahces: RAE+: Configuration as in paper CDLH: Not public, results from paper	
	Response variable, elaboration and metric	Yes	Precision, recall, F1	
Design	Design type	No	Seems 1 factor-4 treatment (RAE+,CDLH,PGD+GGNN, ASTNN)	
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Assume 1	
Instrumentation	Test set	Yes	OJ, BCB, referenced	
	Measuring instruments	Yes		
	Measurement procedure	Yes		
	Technological infrastructure	Same as previous?		
Population	Objects (chars. of the experimental dataset)	Partially	Mention the datasets and references (OJ, BCB) OJ seems to be different from the one used in E1	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified	
Artifact	Availability			
	Badge	Yes	Available badge	

Paper	A novel neural source code represent
ID	AP8
Experiments	4
Comments	

Aspect	Element	E3	Comments	Code
Experiment type		Optimization	Several architectural choices	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		**Same as previous**
	Model parameters	Same as previous?		**Same as previous**
	DL algorithm	Same as previous?		**Same as previous**
	Training hyperparameters	Same as previous?		**Same as previous**
	Training data	Yes	OJ, BCB, referenced	
Operationalization	Factors and treatments	Partially	AST-full/block/node Removing pooling I/II LSMT instead of GRU long code fragments ASTNN	
	Response variable, elaboration and metric	Yes	Accuracy, F1	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Assume 1	
Instrumentation	Test set	Yes	OJ, BCB, referenced	
	Measuring instruments	Yes		
	Measurement procedure	Yes		
	Technological infrastructure	Same as previous?		
Population	Objects (chars. of the experimental dataset)	Same as previous?		
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified	
Artifact	Availability			
	Badge	Yes	Available badge	

Paper	A novel neural source code represent
ID	AP8
Experiments	4
Comments	

Aspect	Element	E4	Comments	Code
Experiment type		Optimization	Batching algorithm	
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		Same as previous
	Model parameters	Same as previous?		Same as previous
	DL algorithm	Same as previous?		Same as previous
	Training hyperparameters	Same as previous?		Same as previous
	Training data	No	Not clear which ones are used	
Operationalization	Factors and treatments	Partially	without batching batching recurrent layer batching recurrent+enconding layers	
	Response variable, elaboration and metric	Yes	Time	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Assume 1	
Instrumentation	Test set	No	Not clear which ones are used	
	Measuring instruments	Yes		
	Measurement procedure	Yes		
	Technological infrastructure	Same as previous?		
Population	Objects (chars. of the experimental dataset)	No	Not clear the ones used	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only 3 threats are listed, not classified	
Artifact	Availability			
	Badge	Yes	Available badge	

Paper DeepPerf: performance prediction for configurable software with deep sparse neural network
ID AP10
Experiments 4
Comments E1 are a series of experiments. Difficult to assess how many, as they are described at a very high level

Aspect	Element	E1	Comments	Code
Experiment type		Optimization	Hyperparameters tuning. Could be several experiments. Very bad described	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: n+2 #neurons/layer: l1 (), l2..n+1(128?), ln+2(1) connections: activation functions: params. Initialization: Xavier (weights)		Matches and completes
	Model parameters	biases: No weights: No		
	DL algorithm	representation: Yes model type: Yes loss function: Yes (MSE) regularization: L1 (only in l2). Lambda, grid search with 30 points logarithmically spaced in 0.01-1000 optimization: Adam (default Tensorflow values), gradient clipping		Matches
	Training hyperparameters	train-test split: learning rate: initial between 0.0001-0.1, dropped by 0.001 #iterations: batch size: Size of training data #epochs: 2000?		Matches and completes learning rate: initial between 0.0001-0.1, dropped by 0.001 #epochs: 2000 Train-test split calculated
	Training data	Yes	Input and output are normalized (0-1 and 0-100). Explicitly linked to artifact in paper	
Operationalization	Factors and treatments	Partially	Could be several experiments. Not sure if all hyperparams are made explicit Factors (at least): regularization, #hidden layers, learning rate No levels given	Info is in the code
	Response variable, elaboration and metric	No	time (?)	Testing error
Design	Design type	No		
	Blocking variables	No	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	Partially	#neurons/layer, #epochs, but no value given	
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Input normalized (0-1). Explicitly linked to artifact in paper	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Python 3.6, Tensorflow 1.8.0		
Population	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other publications where they are fully explained	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	E1 are a series of experiments. Difficu

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1, RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Best from previous?		Same as previous
	Model parameters	Best from previous?		
	DL algorithm	Best from previous?		Same as previous
	Training hyperparameters	Best from previous?		Same as previous
	Training data	Yes	Input and output are normalized (0-1 and 0-100)	Same as previous
Operationalization	Factors and treatments	Model Type, Subject system (?)	DECART, DeepPerf (DECART is classification trees) Not sure if subject system (apache, x264,BDB-J, LLVM, BDB-C, SQLite). Could be blocking variable	Same as previous
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time	
Design	Design type	No		
	Blocking variables	Yes	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Input normalized (0-1). Explicitly linked to artifact in paper	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Python 3.6, Tensorflow 1.8.0		
Population	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other publications where they are fully explained	
Analysis	Descriptive statistics	Yes	Mean and 95% CI	
	Inferential statistics	Yes	t-test	
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	E1 are a series of experiments. Difficu

Aspect	Element	E3	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ2, RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Best from previous?		Same as previous
	Model parameters	Best from previous?		
	DL algorithm	Best from previous?		Same as previous
	Training hyperparameters	Best from previous?		Same as previous
	Training data	Yes	Input and output are normalized (0-1 and 0-100)	Same as previous
Operationalization	Factors and treatments	DNN architecture	SPLConqueror (no DNN), DeepPerf (DECART is classification trees)	Same as previous
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time	
Design	Design type	No		
	Blocking variables	Yes	N-fold validation (30 times resampling training set)	It seems this could be changed in the call
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Input normalized (0-1). Explicitly linked to artifact in paper	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Python 3.6, Tensorflow 1.8.0		
Population	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other publications where they are fully explained	
Analysis	Descriptive statistics	Yes	Mean and 95% CI	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	DeepPerf: performance prediction for
ID	AP10
Experiments	4
Comments	E1 are a series of experiments. Difficu

Aspect	Element	E4	Comments	Code
Experiment type		Optimization	Different architectures (SVM, dropout, L1, L2, no regularization)	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Best from previous?		
	Model parameters	Best from previous?		
	DL algorithm	Best from previous?		
	Training hyperparameters	Best from previous?		
	Training data	Yes	Input and output are normalized (0-1 and 0-100)	
Operationalization	Factors and treatments	DNN architecture	DeepPerf, L1-all-FNN, Plain-FNN, L2-FNN, Dropout-FNN Also SVM, but it is not a DNN For the others mentions some hyperparameters, but they are not fully described	
	Response variable, elaboration and metric	Yes	Mean Relative Error (MRE), training time	
Design	Design type	No		
	Blocking variables	Yes	N-fold validation (30 times resampling training set)	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Input normalized (0-1). Explicitly linked to artifact in paper	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Python 3.6, Tensorflow 1.8.0		
Population	Objects (chars. of the experimental datasets)	Partially	Briefly describes them. References are given to other publications where they are fully explained	
Analysis	Descriptive statistics	Yes	Mean and 95%CI	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Internal, external	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper Detection of hidden feature requests from massive chat messages via deep siamese network
ID AP15
Experiments 3
Comments Not sure the extent of the proposed solution. Mention data preparation, and preprocessing is one step

Aspect	Element	E1	Comments	Code
Experiment type		Optimization+Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 2+4(x2)+1+1 #neurons/layer: l1, l2 (512) connections: l1, l2 (feedforward), l3 (input), l4 (convolutional), l5 (BiLSTM), l6 (combination), l7 (similarity) activation functions: l1, l2 softsign params. Initialization: l1, l2 (trained)	2 first layers are SOTA disentanglement (already trained). Not clear if this is approach or not. I would say not, as it is not trained. All descriptions are partial. Grid search used for: POS tag embedding (50), kernel sizes (2,3,4,5) feature maps/kernel (25), output dimension of BiLSTM is 300 (150 for each direction). I assume these are the "chosen" but do not know the initial ones	**Now fully addressed. No contradictory info** e.g. activation functions: l4 relu, l7 sigmoid
	Model parameters	biases: No weights: No	Although l1 and l2 corresponds to a SOTA disentanglement NN, and they are available	
	DL algorithm	representation: Partially model type: Partially loss function: Cross-entropy regularization: Dropout (0.1) and early stopping (after 10 epochs) optimization: No	It is not clear if their approach includes disentanglement or not. They mention dropout and early stopping as optimization, but it seems to me they are regularization	**loss function: Cross-entropy. But also contrastive, MSE and logits appear in frminer_model.py** **optimization: dense_sparse_Adam in config.json, Adam in finetune_config.json** **regularization: dropout (0.1) input embeddings & similarity layer**
	Training hyperparameters	train-test split: Yes learning rate: No #iterations: No batch size: No #epochs: No	3-fold intra-project-cross-validation from 3 projects	learning rate: 1e-4 **#epochs: 10 in config.json, 80 in finetune_config.json, install.MD 100** batch size: 32 **No reference in the code to grid search**
	Training data	Yes	Explicitly lined to artifact	**No train_d and train_t**
Operationalization	Factors and treatments	DNN model	FRMiner,, p-FRMiner, CNC, FT Others not explained (only p-FR miner). For CNC, codes and models provided in the publication. For FT, official released packages, trained (100 epochs, initial learning rate 1.0, n-gram 2), and hyperparameters tuning .	
	Response variable, elaboration and metric	Yes	Precision, recall, F1	
Design	Design type	No		
	Blocking variables	No	Project?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	Yes	Cross-validation	
Instrumentation	Test set	Yes	Same as training data	**No train_d and train_t**
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Partially	Allennlp (open-source NLP library built on PyTorch). Missing versions NVIDIA 1060 GPU, intel core i7, 16GB RAM, Ubuntu	
Population	Objects (chars. of the experimental datasets)	No	A table with some info is given, but nothing is said	
Analysis	Descriptive statistics	Partially	Average reported from 3-fold-cross-val	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	External, internal, construct (the authors define it for RV only)	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Detection of hidden feature requests f
ID	AP15
Experiments	3
Comments	Not sure the extent of the proposed s

Aspect	Element	E2	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?	Same for FRminer and p-Frminer	Same as previous
	Model parameters	Same as previous?		
	DL algorithm	Same as previous?	Same for FRminer and p-Frminer	Same as previous
	Training hyperparameters	Same as previous	Same for FRminer and p-Frminer	Same as previous
	Training data	Same as previous?	Same for FRminer and p-Frminer	Same as previous
Operationalization	Factors and treatments	DNN model, dataset size	FRMiner, p-FRMiner (initial, x5, x10, x20, x30)	
	Response variable, elaboration and metric	Yes	Precision, recall, F1	
Design	Design type	No		
	Blocking variables	No	Project?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	Yes	Cross-validation	
Instrumentation	Test set	Yes	Same as training data	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Same as previous		
Analysis	Descriptive statistics	Partially	Average reported from 3-fold-cross-val	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	External, internal, construct (the authors define it for RV only)	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Detection of hidden feature requests f
ID	AP15
Experiments	3
Comments	Not sure the extent of the proposed s

Aspect	Element	E3	Comments	Code
Experiment type		Optimization+Evaluation		
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous?		Same as previous
	Model parameters	Same as previous?		
	DL algorithm	Same as previous?		Same as previous
	Training hyperparameters	Same as previous	3-fold cross-project-cross-validation from 3 projects	Same as previous
	Training data	Same as previous?		Same as previous
Operationalization	Factors and treatments	DNN model	Same as E1	
	Response variable, elaboration and metric	Yes	Precision, recall, F1	
Design	Design type	No		
	Blocking variables	No	Project?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	Yes	Cross-validation	
Instrumentation	Test set	Yes	Same as training data	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Same as previous		
Analysis	Descriptive statistics	Partially	Average reported from 3-fold-cross-val	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	External, internal, construct (the authors define it for RV only)	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	On using machine learning to identify knowldge in API reference documentation
ID	AP29
Experiments	1
Comments	

Aspect	Element	E1	Comments	Code
Experiment type		Comparison	SOTA	
Hypotheses	Research hypotheses	Yes	RQ1, RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 5 #neurons/layer: 300, ?, 128, 64, 12 connections: input, LSTM, Dense, Dense, output activation functions: tahn,, ReLU, ReLU, sigmoid params. Initialization: No		#layers: 5 #neurons/layer: **50**, **256**, 128, 64, 12 connections: input, LSTM, Dense, Dense, output activation functions: tahn,, ReLU, ReLU, sigmoid params. Initialization: Default
	Model parameters	biases: No weights: No		
	DL algorithm	representation: Yes model type: Yes loss function: sigmoidal cross-entropy regularization: No optimization: dropout (en LSTM), Adam	Input word embedding vectors trained using GloVe	representation: Yes model type: Yes loss function: **binary** cross-entropy regularization: No optimization: dropout (en LSTM=**0.2**), Adam
	Training hyperparameters	train-test split: Yes learning rate: 0.001 #iterations: No batch size: 32 #epochs: 100	10-fold cross-validation using 10% of dataset as test set	train-test split: v**alidation=0.2, test=0.1** learning rate: 0.001 #iterations: No batch size: 32 #epochs: 100 **Early stopping**
	Training data	Yes (paper mentions code and data of the study are shared	CADO. Resampling is made to improve it	**NO. Must be requested from authors**
Operationalization	Factors and treatments	Partially (at different levels)	Two algorithms (k-NN and SV), and RNN with LSTM layer architecture, naïve (MF1, MF2, RAND))	**Implements RNN, k-NN and SV. But in readme, d-NN and SV are not mentioned to be run. Additionally, k-NN code needs CADO. SV needs training data not provided**
	Response variable, elaboration and metric	Partially	Not all are described at the same level of detail AUPRC (per knowledge type), hamming loss, subset accuracy, macroprecision, macrorecall, macroF1, macroAUC	**NO** **I can only see for RNN F1 and 'accuracy'**
Design	Design type	No		
	Blocking variables	Corpora used to train embeddings Knowledge type Test set	Glove is trained on 4 corpora for RNN	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	10	10-fold cross-validation using 10% of dataset as test set	10
Instrumentation	Test set	Partially	New Python dataset (not very well explained)	**Available**
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Partially	GloVe for embeddings, trained on 4 corpora	
Population	Objects (chars. of the experimental datasets)	Partially	New Python dataset (not very well explained)	
Analysis	Descriptive statistics	Partially	10-fold cross-validation, assume using means	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Not linked in al cases to classification	

Paper	MTFuzz: Fuzzing with a multi-task neural network
ID	AP36
Experiments	4
Comments	

Aspect	Element	E1	Comments	Code
Experiment type		Comparison		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: 7 + 3(with 3 paralells) #neurons/layer: L1(?), L2(2048), L3(1024), L4(512) connections: 3 encoder, 3 (x3) decoder activation functions: ReLu for hidden, sigmoid output params. Initialization:	Each task has the same weight	#layers: 7 + 3(with 3 paralells) #neurons/layer: L1(**2048**), L2(2048), L3(1024), L4(512), **for decoder?: 8,32 (commented),1** connections: 3 encoder, 3 (x3) decoder activation functions: ReLu for hidden, sigmoid output **params. Initialization: random**
	Model parameters	biases: No weights: No		Model is saved and reloaded, but not stored. For avoiding crashes
	DL algorithm	representation: Yes model type: Yes loss function: multi-task regularization: No optimization: Adam	Loss function: MSE for edge coverage, adaptive loss for edge and context-sensitive edge	representation: Yes model type: Yes loss function: multi-task **regularization: No** optimization: Adam
	Training hyperparameters	train-test split: No learning rate: 0.001 #iterations: No batch size: No #epochs: 100	750 input samples for re-training	**train-test split: Yes** learning rate: 0.001, **also 0.1 for decoder?** **#iterations: 50 (with 100 epochs)** **batch size: Yes (with 300 epochs)** #epochs: 100. **But also allows 300**
	Training data	Yes		**Names do not match the ones in the paper**
Operationalization	Factors and treatments	Partially	Fuzzer: AFL, AFLFasst, FairFuzz, Angora (non-DNNs), Neuzz(DNN), MTFuzz	
	Response variable, elaboration and metric	Yes	Number of bugs detected, edge coverage	
Design	Design type	No		
	Blocking variables	Program	10 programs	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	Yes	24 hours for real-world, 5 hours for synthetic bugs	
	Procedure	No		
	Number of experimental units	Yes	5 repetitions to cover fuzzer variability	
Instrumentation	Test set	Yes	2 datasets, one for real bugs, other for synthetic	**Names do not match the ones in the paper**
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu18.04, Intel Xeon E5-2623, NVIDIA GTX 1080Ti GPU. For data collection, single core machine for an hour	
Population	Objects (chars. of the experimental datasets)	Partially	Nothing for synthetic bugs	
Analysis	Descriptive statistics	Partially	For edge coverage mean and std. Dev.	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

Paper MTFuzz: Fuzzing with a multi-task neu
ID AP36
Experiments 4
Comments

Aspect	Element	E2	Comments	Code
Experiment type		Optimization	With some/without auxiliary tasks	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		**Same as previous**
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		**Same as previous**
	Training hyperparameters	Same as previous		**Same as previous**
	Training data	Same as previous		**Same as previous**
Operationalization	Factors and treatments	Partially	Removing some of the decoders. All configs. Use same hyperparams, etc.	
	Response variable, elaboration and metric	Yes	Edge coverage	
Design	Design type	No		
	Blocking variables	Program		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	Yes	1 hour	
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Real bugs only	**Names do not match the ones in the paper**
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu18.04, Intel Xeon E5-2623, NVIDIA GTX 1080Ti GPU. For data collection, single core machine for an hour	
Population	Objects (chars. of the experimental datasets)	Partially		
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

Paper	MTFuzz: Fuzzing with a multi-task ne
ID	AP36
Experiments	4
Comments	

Aspect	Element	E3	Comments	Code
Experiment type		Optimization	Adaptive loss	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		Same as previous
Operationalization	Factors and treatments	Yes	Adaptive loss	
	Response variable, elaboration and metric	No	Recall, F1	Completes
Design	Design type	No		
	Blocking variables	Program		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No	Not specified this time	
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Real bugs only	Names do not match the ones in the paper
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu18.04, Intel Xeon E5-2623, NVIDIA GTX 1080Ti GPU. For data collection, single core machine for an hour	
Population	Objects (chars. of the experimental datasets)	Partially		
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

Paper	MTFuzz: Fuzzing with a multi-task ne
ID	AP36
Experiments	4
Comments	

Aspect	Element	E4	Comments	Code
Experiment type		Generalization		
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		Same as previous
Operationalization	Factors and treatments	Yes	Program type: ELF files, XML files, Fuzzer (Neuzz, MTFuzz, AFL, MTFuzz inputs+embeddings)	
	Response variable, elaboration and metric	Yes	Edge coverage	
Design	Design type	No		
	Blocking variables	Program	3 ELF, 2 XML	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	Yes	1 hour	
	Procedure			
	Number of experimental units		Seems 1 run	
Instrumentation	Test set	Partially	Only reference	Names do not match the ones in the paper
	Measuring instruments	Deduced		
	Measurement procedure	Deduced		
	Technological infrastructure	Yes	Keras 2.2.3 with Tensorflow-1.8.0. Ubuntu18.04, Intel Xeon E5-2623, NVIDIA GTX 1080Ti GPU. For data collection, single core machine for an hour	
Population	Objects (chars. of the experimental datasets)	Partially		
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	3 threats not classified	
Artifact	Availability	Yes		
	Badge	Yes		

Paper	STATEFORMER: Fine-grained type recovery from binaries using generative state modeling
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary material only. E12 is referenced in supplementary material

Aspect	Element	E1	Comments	Code
Experiment type		Evaluation	Compares against nothing	
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: Some #neurons/layer: No connections: Yes activation functions: No params. Initialization: No	Architecture is described at a high level. For details points to supplementary material	Complete in artifact. But possible incosistencies between supplementary material and code
	Model parameters	biases: No weights: No	Does not explicitly mention parameters	Pre-trained available in artifact, but broken link
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: No optimization: No	Loss: MSE + BCE. Points to supplementary material	In code
	Training hyperparameters	train-test split: 80-10-10 learning rate: No #iterations: No batch size: No #epochs: 10, 50	Pretrain+train Points to supplementary material	In code
	Training data	Yes	Details in supplementary material	There are 3 sets available in dropbox: raw data, preprocessed pre-training and training (fine-tuning). The links of the last 2 are broken. The organization of the first one does not seem to match the paper, and the datasets are not the same as the ones described in the supplementary material. Some are missing, and the others seem to have several versions (not the latests as stated in the paper)
Operationalization	Factors and treatments	Model type (STATEFORMER)	STATEFORMER performance is evaluated	
	Response variable, elaboration and metric	Yes	Precision, Recall , F1	
Design	Design type	No		
	Blocking variables	Deduced	Architecture/optimization/obfuscation	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Details in supplementary material	There are 2 sets available in dropbox: raw data and preprocessed pre-training. The link of the later is broken. The organization of the former does not seem to match the paper, and the datasets are not the same as the ones described in the supplementary material. Some are missing, and the others seem to have several versions (not the latests as stated in the paper)
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Yes	Pytorch 1.6.0 (Fairseq toolkit) Linux server w Ubuntu 18.04 Intel Xeon 4212 2.2.0GHz 48 virtual cores 188GB RAM 4 Nvidia RTX 2080-Ti GPUs pyelftools, Ghidra	
Population	Objects (chars. of the experimental datasets)	Yes	Details in supplementary material	Yes
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Model type (STATEFORMER, EKLAVIA)	For EKLAVIA, numbers reported in paper are used.	
	Response variable, elaboration and metric	Yes	Accuracy	
Design	Design type	No		
	Blocking variables	Deduced	Architecture/optimization	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same 8 projects as EKLAVIA	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	EKLAVIA projects. Supplementary material	Same as previous
Analysis	Descriptive statistics	Partially	Average	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E3	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Model type (STATEFORMER, Debin)	Debin already trained model is used. STATEFORMER is restricted to only 17 types, as Debin	
	Response variable, elaboration and metric	Yes	F1	
Design	Design type	No		
	Blocking variables	Deduced	Architecture/optimization	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	OpenSSL	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	Partially	Average	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E4	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Model type (STATEFORMER, Typeminer)	Typeminer is not open-source. Authors are contacted and asked for the numbers. It is not DNN	
	Response variable, elaboration and metric	Yes	F1	
Design	Design type	No		
	Blocking variables	Deduced (Task)	1 architecture 1 optimization. The ones used by Typeminer 4 Tasks	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Mention evaluated on "their" projects	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	Partially	Average	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E5	Comments	Code
Experiment type		Evaluation	Against SOTA	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Model type (STATEFORMER, Debin, Ghidra)	Ghidra is commercial tool (not DNN)	
	Response variable, elaboration and metric	Yes	Execution time (seconds)	
Design	Design type	No		
	Blocking variables	Project	4 projects	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Only name the projects	Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	Partially	Average	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E6	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Use of pre-training, masking	Not sure the value of the other once one of them is fixed	Artifact does not seem to contain ablation studies (only the final model)
	Response variable, elaboration and metric	Yes	F1	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	No		Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E7	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Assesses STATEFORMER only		
	Response variable, elaboration and metric	Yes (pre-training loss)	MSE, BCE	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	No		Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E8	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		Same as previous
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous???	Not clear	Same as previous
Operationalization	Factors and treatments	Pre-training (no, STATEFORMER, TREX)	TREX is DNN, but they do not mention where they take it from	
	Response variable, elaboration and metric	Yes	F1	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	No		Same as previous
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	Same as previous		
Population	Objects (chars. of the experimental datasets)	Yes	Supplementary. Material	Same as previous
Analysis	Descriptive statistics	Partially	Average	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Flat list of 3 threats	
Artifact	Availability	Yes		
	Badge	Yes	Available, reusable	

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E9-supplementary	Comments
Experiment type		Assessment (different types predicted)	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
	Model parameters		Same as previous
	DL algorithm		Same as previous
	Training hyperparameters		Same as previous
	Training data		Same as previous
Operationalization	Factors and treatments	Yes (type)	Assesses STATEFORMER on predicting different types
	Response variable, elaboration and metric	Partially	Accuracy, not defined
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	Seems 1 run
Instrumentation	Test set	Yes	Same as previous
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	Deduced from paper.
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	Partially	Average only
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E10-supplementary	Comments
Experiment type		Generalization (cross-project)	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
	Model parameters		Same as previous
	DL algorithm		Same as previous
	Training hyperparameters		Same as previous
	Training data		Same as previous
Operationalization	Factors and treatments	Test set	
	Response variable, elaboration and metric	F1	
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	
Instrumentation	Test set	Yes	Same as previous
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	Deduced from paper
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E11-supplementary	Comments
Experiment type		Comparison	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
	Model parameters		Same as previous
	DL algorithm		Same as previous
	Training hyperparameters		Same as previous
	Training data		Same as previous
Operationalization	Factors and treatments	Model type (STATEFORMER, Debin)	
	Response variable, elaboration and metric	F1	
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	
Instrumentation	Test set	Obfuscated ones	Same as previous
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

Paper	STATEFORMER: Fine-grained type rec
ID	AP39
Experiments	12
Comments	E9-11 appear in supplementary mater

Aspect	Element	E12-supplementary	Comments
Experiment type		Optimization	
Hypotheses	Research hypotheses	No	
	Statistical hypotheses	No	
Variables selection	Model hyperparameters		Same as previous
	Model parameters		Same as previous
	DL algorithm		Same as previous
	Training hyperparameters		Same as previous
	Training data		Same as previous
Operationalization	Factors and treatments	Numerical values embedding, number of layers, layers dimensions,	
	Response variable, elaboration and metric	Training loss	
Design	Design type	No	
	Blocking variables	No	
	Held-constant variables	No	
	Measured variables (covariates)	No	
	Randomization	No	
	Task duration	No	
	Procedure	No	
	Number of experimental units	No	
Instrumentation	Test set	Yes	Same as previous
	Measuring instruments	No	
	Measurement procedure	No	
	Technological infrastructure	Same as previous	
Population	Objects (chars. of the experimental datasets)	Yes	
Analysis	Descriptive statistics	No	
	Inferential statistics	No	
Validity evaluation	Conclusion, internal, construct, external	No	
Artifact	Availability	Yes	
	Badge	Yes	Available, reusable

Paper	A syntax-guided edit decoder for neural program repair
ID	AP40
Experiments	5
Comments	

Aspect	Element	E1	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: Not clear #neurons/layer: connections: No activation functions: Some params. Initialization: No		Included in artifact
	Model parameters	biases: No weights: No		
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: Dropout 0.1 optimization: Adam	Loss function: maximize negative log-likelihood of the oracle edit sequence	Adam and AdamW appear in artifact
	Training hyperparameters	train-test split: 80-20 learning rate: 0.0001 #iterations: No batch size: No #epochs: No		Included in artifact
	Training data	Partially	Explanation. Could be reproduced, but it is not linked	Included in artifact
Operationalization	Factors and treatments	Approaches. Factors but not treatments	jGenProg, HDRepair, Nopol, CapGen, SketchFix, FixMiner, SimFix, Tbar, DLFix, PraPR, AVATAR, Recoder	Possible inconsistencies due to DL algorithm
	Response variable, elaboration and metric	Yes	Number of correct patches without perfect fault localization	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	5 hours		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Defects4J v1.2. Described but not explicitly linked to artifact	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No	No characteristics are provided	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper A syntax-guided edit decoder for neural prograr
ID AP40
Experiments 5
Comments

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation		
Hypotheses	Research hypotheses	Yes	RQ1	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		**Same as previous**
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		**Same as previous**
Operationalization	Factors and treatments	SequenceR, CODIT, DLFix, CoCoNuT, TBar, Recoder		**Possible inconsistencies due to DL algorithm**
	Response variable, elaboration and metric	Number of correct patches with perfect fault localization		
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	5 hours		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Defects4J v1.2. Described but not explicitly linked to artifact	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No	No characteristics are provided	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper	A syntax-guided edit decoder for neural prograı
ID	AP40
Experiments	5
Comments	

Aspect	Element	E3	Comments	Code
Experiment type		Optimization		
Hypotheses	Research hypotheses	Yes	RQ2	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		**Same as previous**
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		It is a factor now, but the different datasets do not appear in paper (thus, artifact does not improve)
Operationalization	Factors and treatments	Removing: modify, subtreecopy, insert, placeholder. With eveverything	But testsets are not expected in the code. It cannot be FA	Variation of training datasets do not appear in paper (thus, artifact does not improve paper)
	Response variable, elaboration and metric	Number of correct patches without perfect fault localization		
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	5 hours		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Defects4J v1.2. Described but not explicitly linked to artifact	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No	No characteristics are provided	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper	A syntax-guided edit decoder for neural prograı
ID	AP40
Experiments	5
Comments	

Aspect	Element	E4	Comments	Code
Experiment type		Generalization+Evaluation	Tested in a diferent dataset	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		**Same as previous**
	Training hyperparameters	Same as previous		Same as E1-E2
	Training data	Same as E1-E2		**Same as E1-E2**
Operationalization	Factors and treatments	Tbar, SimFix, Decoder		
	Response variable, elaboration and metric	Number of correct patches without perfect fault localization		
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	5 hours		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Partially	Defects4J v2.0 Described but not explicitly linked to artifact	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No	No characteristics are provided	
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper A syntax-guided edit decoder for neural prograr
ID AP40
Experiments 5
Comments

Aspect	Element	E5	Comments	Code
Experiment type		Optimization	Diferent sizes of training dataset	
Hypotheses	Research hypotheses	No	No associated RQ in paper	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	same as previous		**Same as previous**
	Training hyperparameters	Same as previous		Same as previous
	Training data	Same as previous		
Operationalization	Factors and treatments	Different sizes of training set: 25%, 50%, 75%, 85%, 90%, 93% 96%, 100%	But testsets are not expected in the code. Therefore, we do not know which partitions exactly have been chosen	Subsets not included in code (as expected)
	Response variable, elaboration and metric	Number of correct patches without perfect fault localization		
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	5 hours		
	Procedure	No		
	Number of experimental units	Yes	5 runs	
Instrumentation	Test set	Partially	Defects4J v1.2. Described but not explicitly linked to artifact	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No	No characteristics are provided	
Analysis	Descriptive statistics	Yes	Boxplot	
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partially	Only external and internal	
Artifact	Availability	Yes		
	Badge	Available		

Paper	Lightweight global and local contexts guided method name recommendation with prior knowledge
ID	AP41
Experiments	7
Comments	

Aspect	Element	E1	Comments	Code
Experiment type		Optimization	Hidden	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as next		
	Model parameters	Same as next		
	DL algorithm	Same as next		
	Training hyperparameters	Same as next		
	Training data	No		
Operationalization	Factors and treatments	Yes	Number of tokens from implementation context (5,10,20)	
	Response variable, elaboration and metric	No		
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No		
Instrumentation	Test set	No		
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No		
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper Lightweight global and local contexts guided m
ID AP41
Experiments 7
Comments

Aspect	Element	E2	Comments	Code
Experiment type		Evaluation	Compares against SOTA	
Hypotheses	Research hypotheses	Yes	RQ3	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	#layers: Yes #neurons/layer: No connections: No activation functions: No params. Initialization: No		Code provided
	Model parameters	biases: No weights: No		
	DL algorithm	representation: Yes model type: Yes loss function: Yes regularization: No optimization: No	"Due to page limit, we only briefly introduce this model in the paper, and more details could be referred to the existing work [57]" Loss function:negative log likelihood of the oracle word for that step	optimization: Adagrad regularization: dropout
	Training hyperparameters	train-test split: Yes learning rate: No #iterations: No batch size: No #epochs: No		learning rate: 0.15 iterations: 500,000 batch_size: 120 epochs: inconsistent, 2 values used, both in comments (10 and 5)
	Training data	Yes	References known datasets (Java-small, Java-med, Java-large, Mnire) publicly available	
Operationalization	Factors and treatments	Model type (10 approaches vs Cognac)	For Mnire, they use results reported in paper. Do not mention other approaches (could be the same)	
	Response variable, elaboration and metric	Yes	Precision, Recall, F-score with formulas	
Design	Design type	No		
	Blocking variables	Dataset	Report results per dataset. Could be factor?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training sets. They train and test with the same test.	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Lightweight global and local contexts guided m
ID	AP41
Experiments	7
Comments	

Aspect	Element	E3	Comments	Code
Experiment type		Generalization+Evaluation	Compares agains SOTA for other task (inconsistencies detection	
Hypotheses	Research hypotheses	Yes	RQ4	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	Reference to known dataset (Liu et al), publicly available	
Operationalization	Factors and treatments	Model type (Liu et al, Mnire, Cognac) Class (consistent, inconsistent)	Do not explain any of the others	
	Response variable, elaboration and metric	Yes	Precision, Recall, F-score	
Design	Design type	No		
	Blocking variables	No	Could class be a blocking variable?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training set	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Lightweight global and local contexts guided m
ID	AP41
Experiments	7
Comments	

Aspect	Element	E4	Comments	Code
Experiment type		Optimization	Ablation study for the task in E2	
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	Same as E2	
Operationalization	Factors and treatments	Model (no caller info, no callee info, no prior knowledge, Cognac)	No further details are given	
	Response variable, elaboration and metric	Yes	F-score	
Design	Design type	No		
	Blocking variables	Dataset	Report results per dataset. Could be factor?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training sets. They train and test with the same test.	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Lightweight global and local contexts guided m
ID	AP41
Experiments	7
Comments	

Aspect	Element	E5	Comments	Code
Experiment type		Optimization	Ablation study for the task in E3	
Hypotheses	Research hypotheses	Yes	RQ5	
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	Same as E3	
Operationalization	Factors and treatments	Model (no caller info, no callee info, no prior knowledge, Cognac) Class (consistent/inconsistent)	No further details are given	
	Response variable, elaboration and metric	Yes	F-score, Accuracy	
Design	Design type	No		
	Blocking variables	No	Could class be a blocking variable?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training set	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Lightweight global and local contexts guided m
ID	AP41
Experiments	7
Comments	

Aspect	Element	E6	Comments	Code
Experiment type		Optimization	use of caller/calle info.	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	References known dataset (Mnire), publicly available	
Operationalization	Factors and treatments	Model type (seq2seq mode	No details are given about seq2seq	
	Response variable, elaboration and metric	Yes	F-score	
Design	Design type	No		
	Blocking variables	No		
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training set.	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	

Paper	Lightweight global and local contexts guided m
ID	AP41
Experiments	7
Comments	

Aspect	Element	E7	Comments	Code
Experiment type		Optimization	Similar to E1, but with all tokens	
Hypotheses	Research hypotheses	No		
	Statistical hypotheses	No		
Variables selection	Model hyperparameters	Same as previous		Same as previous
	Model parameters	Same as previous		
	DL algorithm	Same as previous		Same as previous
	Training hyperparameters	Same as previous		Same as previous
	Training data	Yes	References known datasets(Java-small, Java-med, Java-large, Mnire), publicly available	
Operationalization	Factors and treatments	Yes	Tokens (all vs. 10)	
	Response variable, elaboration and metric	Yes	F-score	
Design	Design type	No		
	Blocking variables	Dataset	Report results per dataset. Could be factor?	
	Held-constant variables	No		
	Measured variables (covariates)	No		
	Randomization	No		
	Task duration	No		
	Procedure	No		
	Number of experimental units	No	Seems 1 run	
Instrumentation	Test set	Yes	Same as training sets. They train and test with the same test.	
	Measuring instruments	No		
	Measurement procedure	No		
	Technological infrastructure	No		
Population	Objects (chars. of the experimental datasets)	No		
Analysis	Descriptive statistics	No		
	Inferential statistics	No	Conclusions based on 1 run. At a guess	
Validity evaluation	Conclusion, internal, construct, external	Partial	Flat list of threats	
Artifact	Availability	Yes		
	Badge	Yes	Available	