DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

EVALUATING CONFIDENCE ESTIMATION IN NLU FOR DIALOGUE SYSTEMS

Ranim Khojah

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2022
Supervisor:	Staffan Larsson and Alexander Berman
Examiner:	Eleni Gregoromichelaki
Keywords:	Natural Language Understanding (NLU), Intent ranking, Confidence Calibration.

Abstract

Background: Natural Language Understanding (NLU) is an important component in Dialogue Systems (DS) which makes the utterances of humans understandable by machines. A central aspect of NLU is intent classification. In intent classification, an NLU receives a user utterance, and outputs a list of N ranked hypotheses (an N-best list) of the predicted intent along with a confidence estimation (a real number between 0 and 1) that is assigned to each hypothesis.

Objectives: In this study, we perform an in-depth evaluation of the confidence estimation of 5 NLUs, namely Watson Assistant, Language Understanding Intelligent Service (LUIS), Snips.ai and Rasa in two different configurations (Sklearn and DIET). We measure the calibration on two levels: rank level (results for specific ranks) and model level (aggregated results across ranks), as well as the performance on a model level. *Calibration* here refers to the relation between confidence estimates and true likelihood, i.e. how useful the confidence estimate associated with a certain hypothesis is for assessing its likelihood of being correct.

Methodology: We conduct an exploratory case study on the NLUs. We train the NLUs using a subset of a multi-domain dataset proposed by Liu et al. (2021) on intent classification tasks. We assess the calibration of the NLUs on model- and rank levels using reliability diagrams and correlation coefficient with respect to instance-level accuracy, while we measure the performance through accuracy and F1-score.

Results: The evaluation results show that on a model level, the best calibrated NLU is Rasa-Sklearn and the least calibrated NLU is Snips, while Watson surpasses other NLUs as the best performing NLU and Rasa-Sklearn as the worst performing NLU. The rank-level results resonate with the model-level results. However, on lower ranks, some measures become less informative due to low variation of the confidence estimates.

Conclusion: Our findings convey that when choosing an NLU for a dialogue system, there is a trade-off between calibration and performance, that is, a well-performing NLU is not necessarily well-calibrated, and vice versa. While the chosen metrics of calibration is clearly useful, we also note some limitations and conclude that further investigation is needed to find the optimal metric of calibration. Also, it should be noted that to some extent, our results rest on the assumption that the chosen metrics of calibration is suitable for our purposes.

Preface

On my first day at university, a teacher told me: ”*You’ll know that you’re learning once your brain starts hurting.*”.

Five years later, my brain still hurts. I’ve learned a lot in this thesis and enjoyed every bit of it!

I’d like to thank my patient supervisors: Staffan Larsson and Alexander Berman, for their great guidance and feedback. I’d also like to express my gratitude to my examiner, Eleni Gregoromichelaki for providing detailed feedback to shape the thesis.

Thank you, mama, baba, and ablacım, for supporting me and always being there to listen to me complain.

Contents

1 Introduction 1

1.1 Research Questions 1

1.2 Aims and Contribution 1

2 Background 2

2.1 Related Work 2

2.2 Terminology 3

2.3 Explanatory Example 3

3 Materials and Technical Details 4

3.1 NLU Services and Frameworks 4

3.1.1 Watson Assistant 4

3.1.2 LUIS 5

3.1.3 Snips 5

3.1.4 Rasa Opensource 5

3.2 Dataset 6

4 Methodology 7

4.1 Case Study Setup 7

4.2 Intent Classification 7

4.3 Evaluation of Confidence Estimation 8

4.3.1 Confidence Calibration 8

4.3.2 Performance 9

5 Results and Analysis 10

5.1 Reliability Diagrams 10

5.1.1 Model-level Results 10

5.1.2 Rank-level Results 11

5.2 Spearman’s Correlation Coefficient 12

5.2.1 Model-level Results 12

5.2.2 Rank-level Results 13

5.3 Performance 14

6 Discussion 17

6.1 Validity of Calibration Measures . . . 17
6.2 Interpreting Calibration and Potential Applications . . . 17
6.3 Performance vs. Calibration . . . 18
7 Ethics and Validity Threats . . . 18
7.1 Construct Validity . . . 18
7.2 Internal Validity . . . 19
7.3 External Validity . . . 19
7.4 Data Fallacies . . . 19
8 Conclusion and Future Work . . . 20
References . . . 21
9 Appendix . . . 23
9.1 Rasa Pipelines . . . 23
9.2 Reliability Diagrams with Standard Deviation . . . 23
9.3 Rank-level Ranks vs Spearman's Correlations Plot with Standard Deviation . . . 23
9.4 Results of Rank-level Spearman's Correlation . . . 24
9.5 Post-hoc Analysis: t-test Calculations . . . 25

1 Introduction

Dialogue Systems (DS) have received much attention in academia and industry in recent years. The main goal of work on dialogue systems is to improve the quality of human-computer dialogues by making them more natural. This is achieved in part through the development of the Natural Language Understanding (NLU) component which is responsible for understanding the semantics of user's utterances. Present day NLUs typically apply machine learning models on unstructured data (i.e., the user utterances) to extract features (e.g., keywords, word counts and word embeddings) and predict the intent of the user accordingly (Jung, 2019; Shridhar et al., 2019).

NLU services and frameworks (henceforth NLUs) are widely used by dialogue developers to allow them to create and train NLU models for dialogue systems. The task of choosing an NLU to use may be informed by knowledge about how well-performing and well-calibrated the NLU is in a specific domain or context.

In work on classification in machine learning, calibration is a property that illustrates how estimated confidences reflect real-world probabilities or true likelihood of predictions (Guo et al., 2017). In the context of dialogue systems, well-calibrated NLUs may have an impact on the performance of the DS by providing reliable output to other components in the DS (e.g., Dialogue Manager (DM)), whereas miscalibration can cause problems. Specifically, over-confidence can cause undesired actions from the DS, and under-confidence can cause undesired control questions and clarification questions which disrupt the flow of the dialogue. These disruptions can lead to decreased efficiency and increased distraction and thus result in serious risks in critical applications and domains, such as healthcare and automotive.

1.1 Research Questions

This study addresses methodological questions regarding how to measure calibration, and aims to answer the following research questions:

RQ1: *To what extent are the state-of-the-art NLUs well-calibrated?*

RQ1.1: How well-calibrated are NLUs on model level?

RQ1.2: How well-calibrated are the NLUs on rank level?

RQ2: *How does are state-of-the-art NLUs perform in intent classification tasks?*

RQ3: *To what extent are calibration and performance of an NLU correlated?*

Note that in RQ1, we are interested in comparing calibration across the NLUs, rather than judging the overall calibration e.g., high or low calibration.

1.2 Aims and Contribution

This study's main contribution is **C1)** providing an evaluation of confidence calibration for state-of-the-art NLUs, something that – to the best of our knowledge – hasn't been previously done. **C2)** We propose and test new ways of measuring calibration of all hypotheses rather than only for top hypothesis on rank

and model levels. Finally, we **C3)** make our evaluation scripts publicly available on GitHub[1] along with the dataset we have used to allow the replication of the study and to ease building on it.

Our evaluation aims to help dialogue developers choose an appropriate NLU and also to adapt their dialogue system to specific NLUs. Since our evaluation results describe the reliability of the NLUs in terms of their calibration and performance, dialogue system developers will have more information both about the properties of the NLUs and on how to interpret the output of a given NLU. For example, depending on the degree of calibration in the NLU, grounding behaviours such as positive and negative feedback (indicating understanding or lack thereof) or contextual or interactive disambiguation (clarification requests) can be motivated. If the confidence estimates reflect true likelihood, then if two (or more) hypotheses have similar confidence estimates, this may indicate the presence of an ambiguity in the user input (from the perspective of the NLU, i.e., disregarding the dialogue context) that needs to be resolved. Conversely, if confidence estimates (especially for non-top ranks) do not reflect true likelihood, then even if the top two (or more) hypotheses have similar estimates, this may not be a reliable indication of ambiguity but rather be caused by noise.

2 Background

2.1 Related Work

In prior work, benchmarks and evaluations have been performed to identify the best NLU service in different domains like Software Engineering (Abdellatif et al., 2021), Meteorology (Canonico & De Russis, 2018), Question Answering (QA) (Braun et al., 2017) and others (McTear et al., 2016; Stoyanchev et al., 2016; Kar & Haldar, 2016; Koetter et al., 2018). Generally, these evaluation studies have been conducted to draw the trade-off line between different NLU services in terms of the usability of their user interfaces (Gregori, 2017), the technical features the NLUs provide (e.g., language and device support) (Koetter et al., 2018) and performance as regards identifying the correct intent of a user's utterance (Braun et al., 2017; Liu et al., 2021).

NLU performance is usually evaluated with accuracy or F1 score (Braun et al., 2017), both of which depend only on the top hypothesis returned by the NLU and disregard the associated confidence estimates. For example, an NLU that predicts 3 out of 10 intents incorrectly with high confidence estimates has the same performance of an NLU that predicts 3 out of 10 intents incorrectly with low confidence estimates.

In addition, various methods for visualizing and measuring confidence calibration (the extent to which confidence estimates reflect true likelihoods) have been discussed in work on machine learning for classification tasks. For example, Liu et al. (2021) and Vasudevan et al. (2019) visualize calibration of neural network models through reliability diagrams. Another proposed calibration measure is the correlation coefficient of confidence estimate with respect to F1 score through Spearman's correlation (Dong et al., 2018) and with respect to instance-level accuracy through Pearson's correlation (Vasudevan et al., 2019). Expected Calibration Error (ECE) (Naeini et al., 2015) has also been used to measure the calibration of many models (Guo et al., 2017; Kuleshov et al., 2018). Furthermore, Nixon et al. (2019) extend ECE in deep learning models by looking into the probabilities of all predictions rather than the top one.

In this study, we apply two of the previously proposed calibration assessment methods to NLUs, that is, reliability diagrams and correlation between confidence estimates and instance-level accuracy. On a model level, we follow Nixon et al. (2019) in considering confidence estimates of all hypotheses returned

[1] https://github.com/ranimkhojah/confidence-estimation-benchmark

by an NLU – including hypotheses with lower ranks (2-N) – but using another measure. We also measure calibration on rank level, enabling a more fine-grained analysis.

2.2 Terminology

NLUs are commonly used to perform intent classification and entity recognition tasks. In intent classification, an NLU takes a user utterance as an input, then it parses it into a machine-readable representation in order to return a prediction of the user's intent accordingly (Tur & De Mori, 2011; Wang et al., 2005).

In dialogue systems, intent classification data is used to train an NLU model, that is, user utterances with a corresponding expected intent for each utterance. When using an NLU, an utterance U is fed to the trained NLU model, and the output normally includes the information in Listing 1.

```
{
  'utterance' : 'U' ,
  'top_intent': 'intent_1',
  'intent_ranking' : {
    'intent_1' : 'conf_1', # hypothesis on rank 1
    'intent_2' : 'conf_2', # rank 2
    ... ,
    'intent_N' : 'conf_N'  # rank N
  }
}
```

Listing 1: Abstract example of a JSON response from an NLU when parsing an utterance.

Given an utterance U, the response from an NLU includes the user utterance and a prediction consisting of the top intent and an intent ranking. The intent ranking consists of N-best intent hypotheses with their corresponding confidence estimates. The confidence estimates reflect how confident the NLU model is regarding each hypothesis.

Different NLUs may have different ways of computing these estimates, and they may be slightly different notions of confidence. However, for the purpose of using the estimates in a dialogue system, we are interested in how well they reflect true probabilities. We will note variations in how confidence estimates are computed, but we will not take them into account in our quantitative tests.

2.3 Explanatory Example

We present Figure 1 as a simplified demonstration of how NLUs are used in dialogue systems, using a scenario where a user asks a dialogue system to perform a task within the home domain.

The user interacts and communicate with the dialogue system through a user interface. The user utterance is transferred to an NLU component which uses the utterance as an input to perform intent classification. Next, it returns a prediction with the top intent and the intent ranking that is sent to a Dialogue Manager (DM) which steers the dialogue accordingly. In case of a high estimated confidence for the top hypothesis in the intent ranking, the DM integrates the user's intent, and information is sent to the Natural Language Generation (NLG) unit that generates a response and sends it to the user.

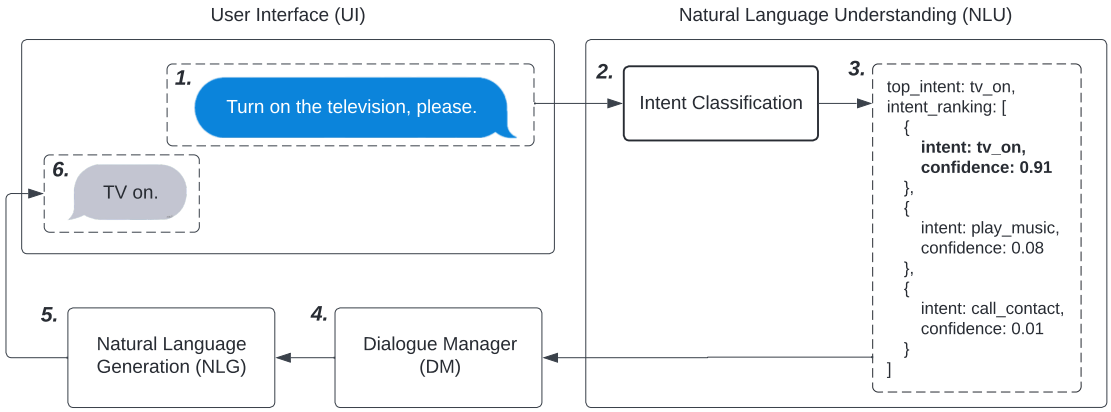


Figure 1: An explanatory example that demonstrates the role of NLU in a dialogue system.

3 Materials and Technical Details

3.1 NLU Services and Frameworks

NLU services and frameworks (henceforth NLUs) can be used to construct the NLU component in a dialogue system. In this study, we choose which NLUs to evaluate based on the following criteria: NLUs that i) can perform intent classification and ii) return at least the 10-best hypotheses in the output. We evaluate 5 commonly-used NLUs: Watson's Assistant (IBM, 2010), Language Understanding Intelligent Service (LUIS) (Microsoft, 2017), Snips (Snips, 2013) and Rasa (Rasa, 2016) with two different pipelines.

In this section, we explain how each of the 5 NLUs is trained and tested through a user interface and/or an API (Summarized in Table 1).

3.1.1 Watson Assistant

Watson Assistant (henceforth Watson) is a cloud-based NLU developed by IBM. It enables managing and building an NLU through the IBM cloud interface which doesn't require any programming experience. Building an NLU includes creating intents and corresponding examples, detecting and resolving conflicts between intents, and training and testing an NLU model. Training an NLU with Watson can only be done through the user interface by manually uploading the training set in a specific format, while testing is possible through the interface and the API.

When parsing an utterance, Watson returns the top 10 hypotheses along with their confidence estimates. Confidence estimates are not normalized as they are calculated independently for each intent that the NLU model has been trained on – in other words, Watson is a multiple-binary classifier which assumes that there is a possibility for an utterance to have more than one correct intent. In addition, Watson has an optional built-in "Irrelevant" intent that corresponds to an out-of-scope (OOS) input[2].

[2]https://cloud.ibm.com/docs/assistant?topic=assistant-irrelevance-detection

3.1.2 LUIS

Language Understanding Intelligent Service (LUIS) is provided by Microsoft and runs on the Azure cloud platform. Similarly to Watson, training is performed via the user interface. Given a training set of intents and their respective user utterances, LUIS trains an intent *I* using its user utterances as positive examples and the user utterances of other intents as negative examples[3] of intent *I*.

There is no limit to the number of hypotheses that LUIS returns; in other words, if the NLU is trained on N intents, the intent ranking is of length N. In the intent ranking, confidence estimates are not normalized, and a "None" intent is included as a representation of an out-of-scope intent. The "None" intent is trained on 0 examples by default, and it requires the user to train it on example utterances[4].

3.1.3 Snips

Snips.ai is an AI voice platform for connected devices (currently acquired by Snonos[5]) which provides an NLU called Snips NLU (henceforth Snips). In this study, we use version v0.20.2 of Snips NLU.

By default, Snips returns all hypotheses of all intents with their confidence estimates that are not normalized, in addition to a built-in "None" intent that is pre-trained by Snips on examples from noise text to cover out-of-scope utterances (Coucke et al., 2018).

3.1.4 Rasa Opensource

Rasa is an open-source NLU provided by Rasa Technologies. It can run on different pipelines that are configurable, which increases the flexibility of the NLU (Bocklisch et al., 2017). The training and testing is performed by sending requests to Rasa Open Source server through Rasa HTTP API[6], or through Rasa NLU's SDK for Python. Rasa returns the top 10 hypotheses by default and their corresponding normalized confidence estimates. In addition, Rasa does not offer a built-in out-of-scope intent.

In this study, we use Rasa v2.4.3 to create two NLU models with two different pipelines (See Appendix 9.1). We use the pre-configured pipelines offered by Rasa. The first pipeline uses an Sklearn Intent Classifier and a pre-trained bag of words (BoW) model where one feature vector is assigned to one input utterance (See Listing 2). In contrast, the second pipeline is built on a Dual Intent and Entity Transformer (DIET) architecture, and is based on a sequence model, meaning that it considers the orders of the words present in an utterance (Bunk et al., 2020). For featurization, this pipeline uses bag of words and n-grams (See Listing 3). We refer to the two pipelines above as Rasa-Sklearn and Rasa-DIET respectively.

NLU	Packaging	Classifier Type	Version/Invoked on	OOS Intent
Watson	Cloud-based service	Multi-binary classifier	Invoked in April 2022	Yes
LUIS	Cloud-based service	Multi-class classifier	Invoked in April 2022	Yes
Snips	Open-source framework	Multi-class classifier	v0.20.2	Yes
Rasa	Open-source framework	Multi-class classifier	v2.4.3	No

Table 1: Summary of NLUs. (*OOS: Out-of-Scope*).

[3]https://docs.microsoft.com/en-us/azure/cognitive-services/luis/luis-concept-model#intents-classify-utterances
[4]https://docs.microsoft.com/en-us/azure/cognitive-services/luis/concepts/intents#none-intent
[5]https://www.sonos.com
[6]https://rasa.com/docs/rasa/http-api/

3.2 Dataset

To conduct intent classification as a part of our evaluation, we use the dataset proposed by Liu et al. (2021). The dataset consists of 25716 annotated user utterances for human-robot interaction and cover 64 Intents, 18 scenarios and 21 domains. The user utterances in the dataset are annotated with, inter alia, the intent, scenario and a normalized version of user utterance which ignores noise, punctuation and converts numbers to text. For instance, the user utterance “Olly, wake me up at 7am!” is annotated with the intent *set* and the scenario *alarm*, and is normalized to ”wake me up at seven am”.

For the NLU evaluation, we generalize the 21 domains and combine them into 6 high-level domains in Table 2. Then we select 10 intents with the most number of examples (See Table 3). Since Watson and Rasa-Sklearn return only the top 10 hypotheses in the intent ranking, by training the NLUs on 10 intents, we ensure consistency across the evaluated NLUs and guarantee that the NLUs will return all the intents that were used to train the NLU. This allows us to include all possible hypotheses in the evaluation.

Domain	**Scenarios**	**Intents**
personal assistance	calendar, datetime, weather, lists, alarm	query, set, remove, convert, createoradd, sendemail, querycontact, addcontact
General	general	quirky, greet, negate, dontcare, repeat, affirm, commandstop, confirm, explain, praise
Audio and vision programmes	music, audio, news, play	query, settings, volume_mute, volume_down, volume_up, music, radio, audiobook, podcasts, game
IoT	iot	hue_lightchange, hue_lightoff, hue_lighton, hue_lightdim, cleaning, hue_lightup, coffee, wemo_on, wemo_off
Outdoor activities	transport, recommendation, social, takeaway	query, ticket, traffic, taxi, locations, events, movies, post, order
Q&A	qa	stock, factoid, definition, maths, currency

Table 2: Domains, scenarios and intents in Liu et al. (2021)’s dataset.

Intent	**Size**	**Normalized Example**
query	5981	what’s the time in australia
set	1748	wake me up at nine am on friday
music	1205	start playing music from favourites
quirky	1088	i am not tired i am actually happy
factoid	1052	tell me comics of charlie chaplin
remove	986	cancel my seven am alarm
negate	939	you don’t understand it right
sendemail	694	send a group mail to lookafter
explain	684	could you clarify me on it further
repeat	585	please let’s start over
Total	14962 examples	

Table 3: Selected intents for the case study with their respective sizes (i.e. number of example utterances) and example utterances from the original dataset.

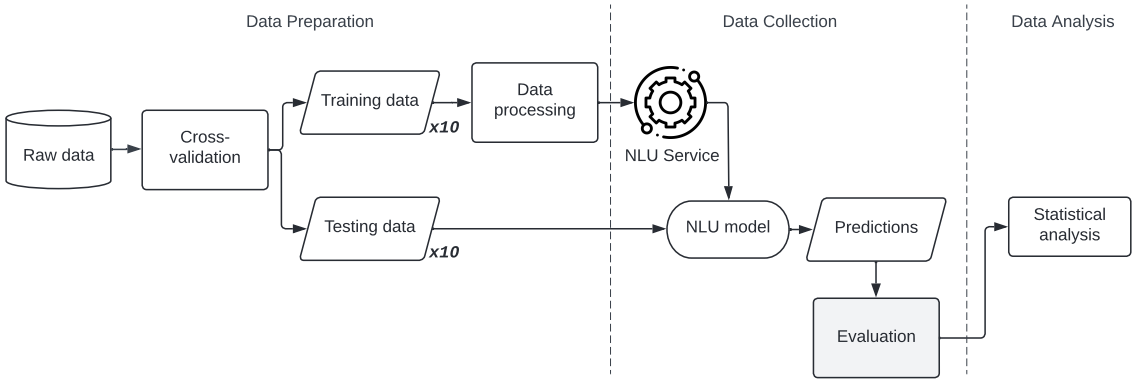


Figure 2: The evaluation process. This process was iterated 5 times, once per NLU. The highlighted box (i.e. Evaluation) is the core component of the study where we assess the confidence estimation of NLUs.

4 Methodology

4.1 Case Study Setup

We conduct an exploratory case study to examine the calibration and performance of the 5 commonly-used NLUs described in Section 3.1 using the multi-domain dataset in Section 3.2. Furthermore, since confidence estimation is the main focus of our study we only evaluate NLUs that return at least the top 10 hypotheses. We summarize the components of our study in Table 4.

An overview of our study's execution is illustrated in Figure 2. We start with selecting the data for the evaluation, applying a cross-validation method (with 10 iterations) and then train an NLU model (See Section 4.2). During data collection, we test the NLU models and obtain results in the form of predictions, that we use to perform the evaluation of the confidence calibration and performance (See Section 4.3). Finally, we verify our evaluation results using a post-hoc statistical analysis (See Section 5).

Objective	**Explore**
The Context	Natural Language Understanding for Dialogue Systems
The Cases	NLUs: IBM Watson Assistant, Microsoft LUIS, Rasa and Snips.ai
Theory	Confidence estimation of NLUs
Research Questions	RQ1, RQ2 and RQ3 (incl. sub-RQs)
NLU Selection Strategy	Cloud-based and open source NLUs that return at least 10 hypotheses
Units of Analysis	Calibration and Performance

Table 4: Case Study Planning (Template from Runeson & Höst (2009)).

4.2 Intent Classification

We select 10 intents with the greatest number of examples (See Table 3) to maximize the size of the dataset and minimize data sparsity. These 10 intents cover 5 domains and 15 scenarios (See Table 2) and have a total of 14962 utterances.

We perform Repeated Random Sub-sampling (Dubitzky et al., 2007) with 10 iterations to generate 10 datasets, where each dataset is split into training and test sets with a 2:1 ratio. Next, the training sets are

processed by first cleaning the example utterances e.g., removing special characters and examples with missing fields.

We train and test 10 NLU models using 10 splits, resulting in 10 trained NLU models for each NLU. In the test results, we ignore hypotheses with "None"/"irrelevant" intent in the intent ranking to ensure that all NLUs have the same intent ranking length and make their results comparable. Alse, we do not normalize the confidence estimates for any of the NLUs (Clarified in Section 7).

4.3 Evaluation of Confidence Estimation

The evaluation of confidence estimation is performed on two levels: rank and model. On rank level, results are obtained for specific prediction ranks. For example, results for rank 1 pertain to the top-ranked (most confident) prediction hypotheses. On model level, the results of all ranks are aggregated for each NLU model.

The main focus of the evaluation is the calibration of the NLUs. However, we also assess performance in order to investigate the correlation between the NLUs' calibration and performance. The former is measured using reliability diagrams and correlation coefficient with respect to instance-level accuracy, and the latter is measured through accuracy and F1-score. Furthermore, the evaluation is conducted for each split and then averaged across splits.

4.3.1 Confidence Calibration

Confidence calibration is the extent to which a model is able to produce confidence estimates that reflect the accuracy (true likelihood) of the respective intent hypotheses (Guo et al., 2017). For example, a model is well-calibrated if hypotheses with confidence estimate 0.7 are correct in 70% of the cases. In this study, we visualize calibration using reliability diagrams and numerically estimate calibration using Spearman's correlation coefficient with respect to instance-level accuracy, as outlined below.

Reliability diagrams: Reliability diagrams are visualizations of a model's calibration (Guo et al., 2017). They plot the true likelihood (accuracy) of a prediction as a function of confidence estimation. Hence, a perfectly-calibrated model is visualized as the identity function, and any deviation indicates miscalibration.

In detail, reliability diagrams are plotted by partitioning predictions into bins, each of which represents a confidence range. In our study, we use 10 uniformly distributed bins, i.e. [0.0-0.1], [0.1-0.2], …[0.9-1.0]. For each bin, mean accuracy is calculated – in other words, the proportion between correct and total number of predictions in the bin (Code in GitHub[7]).

Correlation coefficient with respect to instance-level accuracy: In order to numerically measure the degree of calibration, we compare confidence estimates (scores in the range 0-1) with instance-level accuracies (1 for correct classifications, 0 for incorrect classifications). More specifically, we measure the extent to which an increase in score correlates with an increase in instance-level accuracy – in other words, the monotonicity of the relationship between confidence and accuracy. The degree of monotonicity is measured using the Spearman's correlation coefficient (Xiao et al., 2016)[8].

[7] https://github.com/ranimkhojah/confidence-estimation-benchmark/blob/master/scripts/calibration_evaluation_all_splits.ipynb

[8] We perform Spearman's correlation rather than Pearson's correlation since our data is not normally distributed.

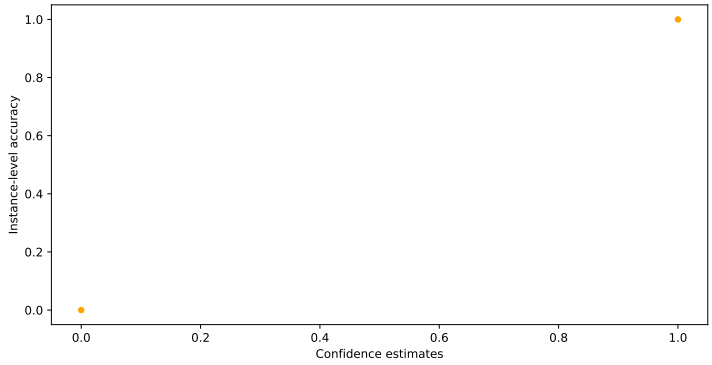


Figure 3: Gold standard confidence estimates: The distribution of confidence estimates with respect to instance-level accuracy of a perfectly-calibrated model (Spearman's correlation of 1.0)

Given two variables (X and Y) of size N each ($x_1, x_2, ..x_n$ and $y_1, y_2...y_n$ respectively), Spearman's correlation coefficient (ρ) is calculated through the formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where n is the number of samples, and d is the pairwise differences of the elements of the variables x_i and y_i.

In intent classification, a perfectly-calibrated NLU has a Spearman's correlation coefficient 1.0, and it always estimates a confidence 1 for correct hypotheses and confidence 0 for incorrect hypotheses, as shown in Figure 3.

Note that other approaches to numerically estimate calibration have been discussed in literature besides Spearman's correlation (Dong et al., 2018), e.g. negative log likelihood and Brier score (Kull et al., 2017) and expected calibration error (Nixon et al., 2019). Different measurement approaches have different advantages and weaknesses (Ashukha et al., 2020; Nixon et al., 2019), and no gold standard seems to exist. In this study, we have opted for Spearman's correlation due to the fact that monotonicity in the relation between confidence and accuracy is an important characteristic of good calibration. Also, it seems intuitive an NLU with more monotonic relation between confidence and accuracy is easier to recalibrate (i.e., improve the calibration) (Nixon et al., 2019).

4.3.2 Performance

Since performance considers only first-ranked hypotheses, it cannot be conducted on a rank level. To measure the performance, we use F1-score and Accuracy. We use F1-score since it considers false positives and false negatives through precision and recall. Another reason is the unbalanced distribution of the example utterances across intents.

However, the complexity of F1-score makes it harder to interpret; therefore, we also include accuracy since in this particular multi-domain dataset, false negatives have no major risks.

Given true positives (TP), true negative (TN), false positives (FP), and false negatives (FN) in the classification results, the following formulas are used to calculate the performance measures.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

5 Results and Analysis

In this section we present our results (averaged across the 10 splits). Our collected data are qualitative in the form of visualizations (reliability diagram) and quantitative (Spearman's correlation, accuracy, F1-score).

For our quantitative results, we provide the average across splits along with standard deviation (SD), whereas for the qualitative results and all visualizations, we plot the averaged results without SD to avoid cluttered diagrams. However, we provide the visualizations with SD of mean accuracies of splits in Appendix 9.2.

Moreover, to verify our quantitative results, post-hoc statistical hypothesis testing is conducted to determine whether there is a statistically significant difference (SSD) between the NLUs' results. We are interested in the SSD between all pairs of NLUs. Therefore, we run a parametric statistical test (t-test) on NLU pairs. Each NLU has 10 rows that represent the results (e.g., Spearman's Correlation, accuracy or F1-score) obtained from each split of the 10 splits. Then, we use Cohen's d to judge the effect size of the statistical significance differences (SSD) between each pair of NLUs and represent the magnitude with the notation: L- *Large*, M- *Moderate*, S- *Small*, N- *Negligible*.

5.1 Reliability Diagrams

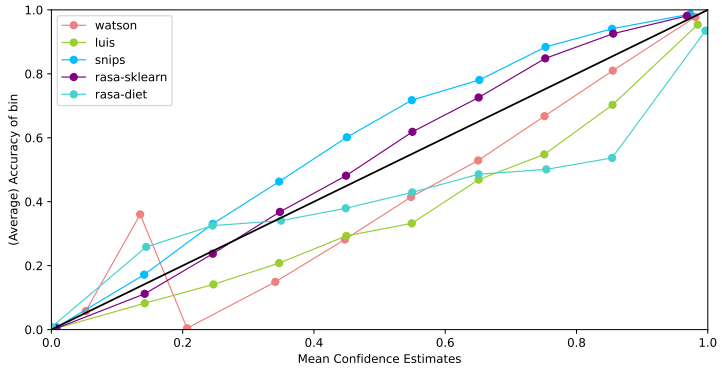
We get a visual overview of the NLUs' calibration through reliability diagrams on a model level (Figure 4) and rank level (Figures 5, 6, 7, 8). In rank-level reliability diagrams, we merge ranks 4-10 due to observed signs of data sparsity; in detail, most of the confidence estimates were within a small range and ended up in 1-2 bins. We also include a histogram with each reliability diagram to visualize the sizes of bins.

5.1.1 Model-level Results

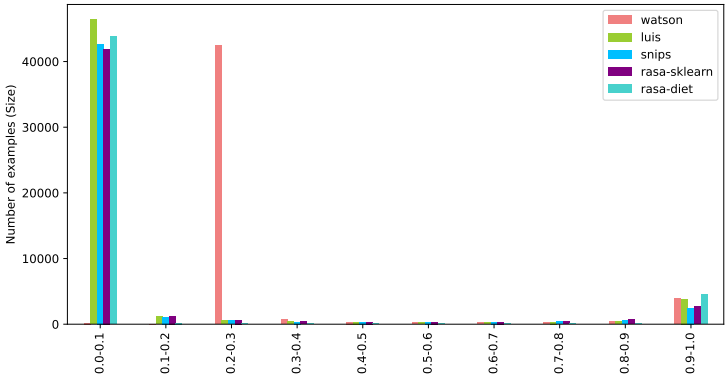
On a model level (Figure 4), all NLUs show a generally monotonous relationship between confidence and accuracy, except for Watson in lower ranges. In general, calibration of NLUs is better in larger bins than smaller bins besides Watson in the range [0.2-0.3] (shown in Figure 4b). In particular, Rasa-Sklearn (the purple curve) is the closest to the gold standard, and is thus the best calibrated NLU according to this analysis. Moreover, Snips underestimates the true likelihood of predictions, while LUIS overestimates predictions.

In contrast, Rasa-DIET's calibration varies depending on the confidence estimate – more specifically, it is both over- and under-confident, for different parts of the confidence range.

Watson underestimates predictions when the confidence estimate is between 0.1 and 0.2, and overestimates otherwise. Furthermore, we observe a discrepancy in Watson's second and third bins in the reliability diagram (Figure 4), a sudden underestimation followed by a drop that indicates an extreme overestimation. According to the bin sizes shown in Figure 4b, Watson's confidences mostly lie within the range 0.2-0.3, whereas the other NLUs' confidence estimates are mostly within the range 0.0-0.1.



(a) Model-level reliability diagram



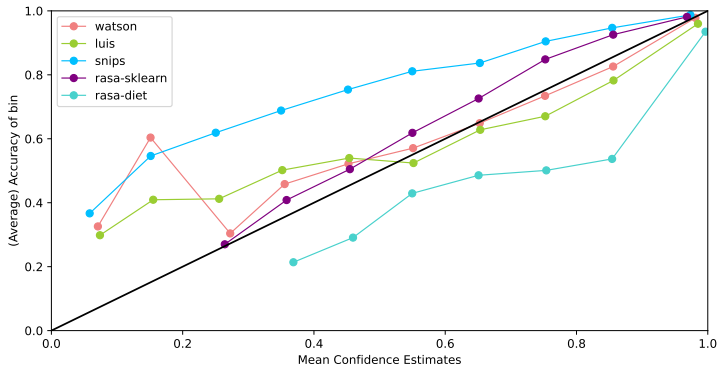
(b) Histogram (Bins sizes)

Figure 4: Model-level reliability diagram (a) and corresponding histogram of bins sizes (b). In the reliability diagram (a), the x-axis shows the mean confidence estimates in 10 bins, while the y-axis shows the mean accuracy of the confidence estimates in each bin (averaged across splits). The black diagonal line plots the identity function representing a gold standard of a perfectly-calibrated model.

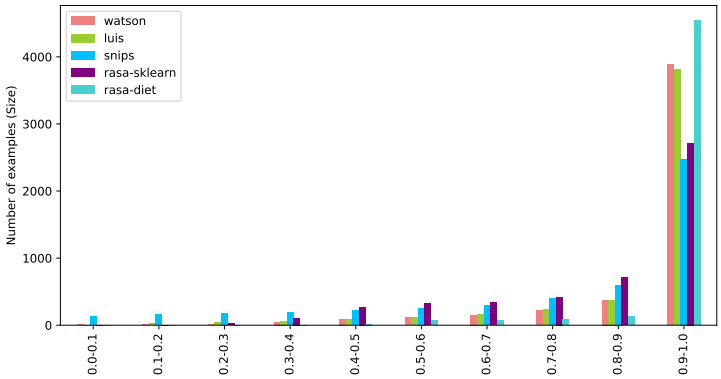
5.1.2 Rank-level Results

On the first rank (Figure 5), the NLUs are fairly well-calibrated in general. Although miscalibration is observed in the first two bins of rank 1 in Watson (confidence estimate between 0.0 and 0.2) which is also confirmed by the high standard deviation in these two bins (See Appendix 9.2).

On ranks 2 (Figure 6) and 3 (Figure 7), the degree of calibration decreases (in comparison with the previous rank), for three of the NLUs (Watson, LUIS and Snips – all over-confident), while for the Rasa NLUs, the trend seems inverted. Moreover, on ranks 4-10 (Figure 8), the reliability diagrams are difficult to interpret due to data sparsity shown in Histogram 8b.



(a) Rank-level reliability diagram (rank 1)



(b) Histogram (Bins sizes)

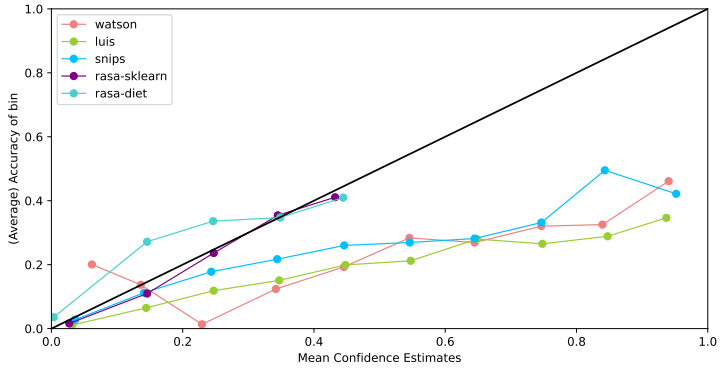
Figure 5: Rank-level reliability diagram (a) and corresponding histogram of bins sizes (b) on rank 1. Note that Rasa-Sklearn and Rasa-DIET don't have any first-ranked hypotheses with a confidence estimate within the first two and three bins respectively. This is due to that Rasa returns normalized confidence estimate whereas the other NLUs don't.

5.2 Spearman's Correlation Coefficient

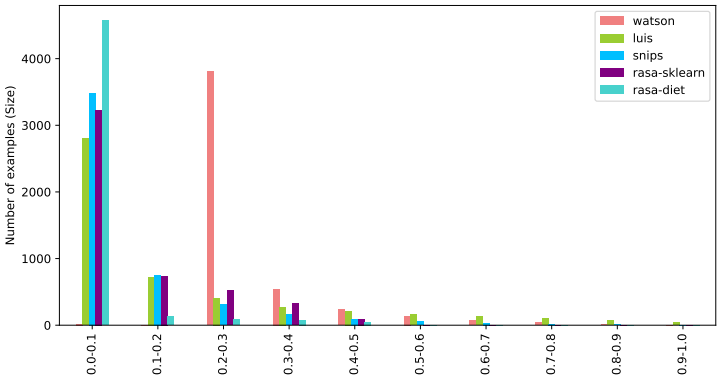
5.2.1 Model-level Results

The calculated Spearman's correlations between the confidence estimates and instance-level accuracy (Table 5) show that Rasa-Sklearn has the highest Spearman's Correlation with a score of $\sim$0.510, and is followed by LUIS, Rasa-DIET, Watson and Snips with the lowest Spearman's correlation of $\sim$0.507. In addition, LUIS and Rasa-DIET are not significantly different, while the differences between each other pair of NLUs is significantly different with a large effect size as detailed in the t-test results in Table 9 in Appendix 9.5.

Since the reliability diagrams show a high degree of monotonicity in the relation between confidence and accuracy, one may expect Spearman's correlations to be close to 1.0. To this end, the actual range of Spearman's correlation coefficients (0.50-0.51) could be seen as an indication of a conflict between the two types of measurement. However, this apparent conflict can be explained by the distribution of the confidence estimates with respect to the instance-level accuracies. Figure 9 plots the confidence estimates of hypotheses returned by Rasa-sklearn. It shows that the confidence estimates assigned to both correct and incorrect hypotheses are within a wide range (i.e., 0.0-1.0). As illustrated by the gold standard in Figure 3, a model can achieve a Spearman's correlation of 1.0 only when it always estimates a



(a) Rank-level reliability diagram (rank 2).



(b) Histogram (Bins sizes)

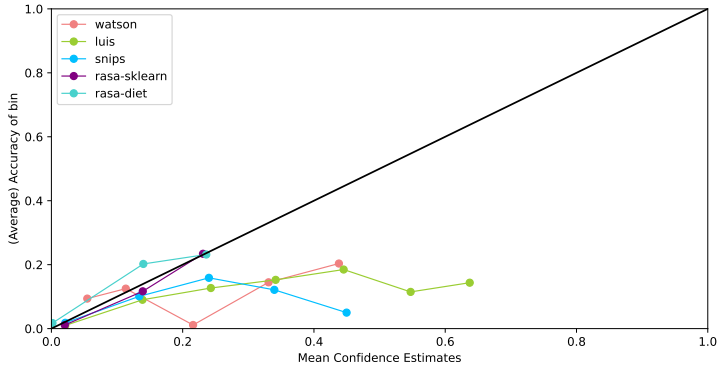
Figure 6: Rank-level reliability diagram (a) and corresponding histogram of bins sizes (b) on rank 2. Note that the ranges of non-empty bins for Rasa-Sklearn and Rasa-DIET become narrower in comparison with rank 1 (Figure 5). In particular, second-ranked hypotheses are not assigned a confidence estimate higher than 0.5.

confidence 1.0 for correct hypotheses and confidence 0.0 for incorrect hypotheses – in other words, when it has perfect accuracy. Hence, in cases with imperfect overall accuracy – e.g. in the presence of genuine ambiguity – a Spearman's correlation of around 0.5 is not necessarily an indication of poor calibration.

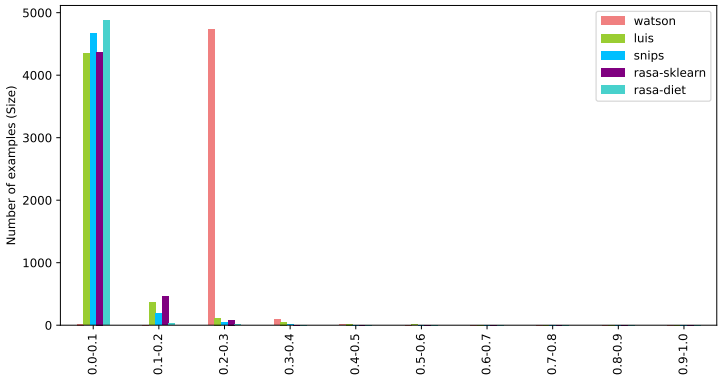
5.2.2 Rank-level Results

As for the **rank-level** Spearman's correlation results in Appendix 9.4, a trend is observed across all ranks. To illustrate the trend, we propose a new kind of diagram in Figure 10 (with SD in Appendix 9.3) that has not been provided in literature, where the NLUs' correlation coefficients start to drop after rank 1 at a decreasing rate, then they level off and reach a minimum on the lowest rank.

Furthermore, the results of each NLU (in Table 8) with respect to the post-hoc analysis (in Table 10) show that the NLU with the significantly highest Spearman's correlations is Rasa-Sklearn (on ranks 1-3) with a large effect size, and LUIS (on ranks 4-7) with a moderate/large effect size. On ranks 8, 9 and 10, the results are either not significant or significant with a small or negligible effect size. In summary, Rasa-Sklearn shows a significantly best calibration on ranks 1-3, while LUIS is significantly best-calibrated on ranks 4-7.



(a) Rank-level reliability diagram (rank 3)



(b) Histogram (Bins sizes)

Figure 7: Rank-level reliability diagram (a) and corresponding histogram of bins sizes (b) on rank 3. Note that the ranges of non-empty bins for Rasa-Sklearn and Rasa-DIET become narrower in comparison with rank 2 (Figure 6). In particular, third-ranked hypotheses are not assigned a confidence estimate higher than 0.3.

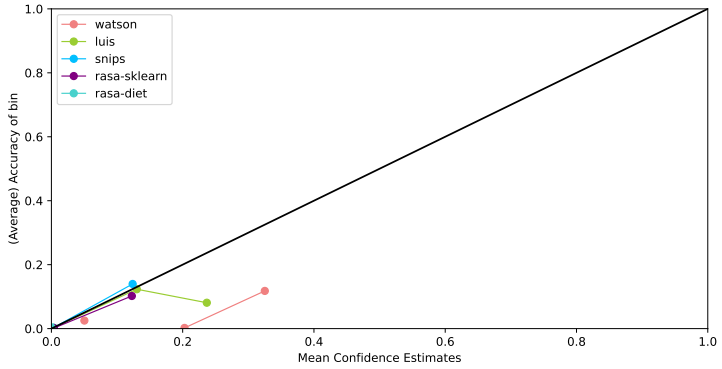
The decrease of Spearman's correlation in lower ranks indicates that the monotonicity drops for all NLUs the lower we go in the intent ranking. In particular, for hypotheses lower than rank 4, the confidence estimates are not informative nor significant. Third-ranked hypotheses are roughly half as monotonous (Spearman's Correlation around 0.2) as top-ranked ones (Spearman's Correlation around 0.4) with second-ranked ones in between (Spearman's Correlation around 0.3). However, there are other factors that can cause such pattern, such as the variation of the confidence estimates (further discussed in Section 6).

5.3 Performance

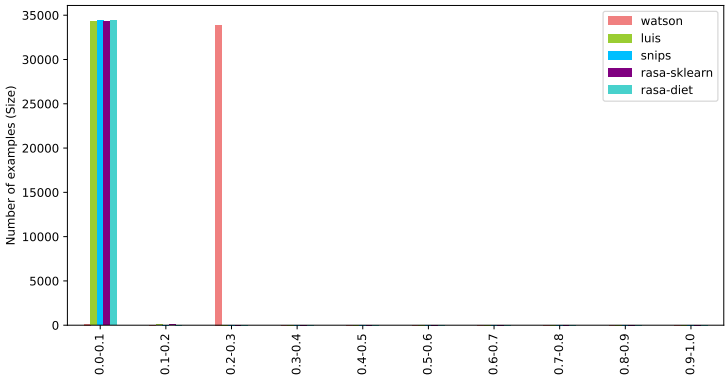
In order to investigate the correlation between the calibration and performance of NLUs, we measure the performance of the NLUs in intent classification by evaluating the accuracy and the F1-score. In this section, we present the averaged results of accuracy and F1-scores across 10 splits for each NLU.

Accuracy: The results in Table 6 show that Watson surpasses all NLUs with ∼0.92 accuracy, followed by Rasa-DIET, Snips, LUIS and Rasa-Sklearn with the lowest accuracy score of ∼0.87.

F1-score: The results of the F1-scores in Table 7 show that Watson achieves the highest F1-score of ∼0.92, followed by Snips, LUIS, Rasa-DIET and Rasa-Sklearn with a score of ∼0.79.



(a) Rank-level reliability diagram (ranks 4-10)



(b) Histogram (Bins sizes)

Figure 8: Rank-level reliability diagram (a) and corresponding histogram of bins sizes (b) on ranks 4-10.

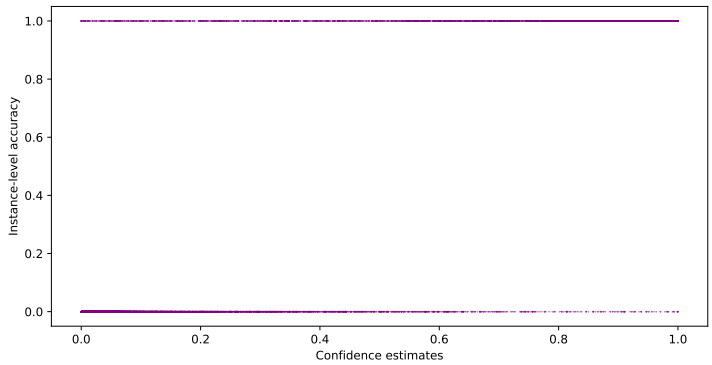


Figure 9: Rasa-Sklearn confidence estimates: The distribution of confidence estimates with respect to instance-level accuracy of Rasa-sklearn which is the best calibrated model across the evaluated NLUs (Spearman's correlation of $\sim$0.51).

Also, according to the analysis in Table 11, the performance scores of LUIS and Snips are not significantly different, meaning that LUIS and Snips can possibly achieve similar results. Also, all of the pairwise performances between the NLUs is significant with a large effect size.

Overall, the performance results are consistent with the results of the following studies that have 3 NLUs

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	0.50838	0.50935	0.50669	**0.51024**	0.50906
Median	0.50851	0.50934	0.506491	**0.51026**	0.50888
***p*-value**	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
SD	0.00075	0.00055	0.00064	0.00046	0.00074

Table 5: Model-level Spearman's Correlation Coefficient ρ: The mean, median, p-value and the standard deviation (SD).

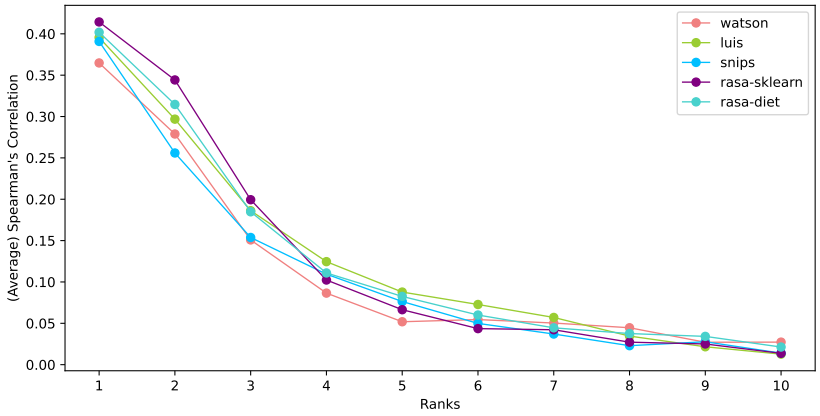


Figure 10: Rank-level Spearman's Correlation for Watson, LUIS, Snips, Rasa-Sklearn and Rasa-DIET on ranks 1-10.

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	**0.92287**	0.88726	0.88991	0.87263	0.90376
Median	**0.91997**	0.890405	0.89060	0.87866	0.89973
SD	0.00225	0.00417	0.00414	0.00386	0.003860

Table 6: (Averaged) Accuracy Scores of NLUs in addition to the median and the standard deviation (SD).

NLU	Watson	LUIS	Snips	Rasa-Sklearn	Rasa-DIET
Mean	**0.92144**	0.88890	0.89029	0.79020	0.81890
Median	**0.91972**	0.89300	0.89166	0.79561	0.81716
SD	0.00234	0.00373	0.00407	0.00358	0.00331

Table 7: (Averaged) F1-scores of NLUs in addition to the median and the standard deviation (SD).

in common (i.e., Watson, LUIS and Rasa-Sklearn): 1) Liu et al. (2021) who use the complete version of our dataset, and 2) Abdellatif et al. (2021) who use two datasets from the software engineering domain. However, it's different from the results in (Braun et al., 2017) that use Telegram chatbot and StackExchange corpora in QA domain; their results show that Watson is the worst performing NLU, and Rasa and LUIS are on top. In addition, Coucke et al. (2018) builds on Braun et al. (2017)'s evaluation by adding Snips. Their results show that Snips outperforms Rasa-Sklearn, as well as Watson using chatbot

and askUbuntu datasets.

6 Discussion

In this section, we focus on interpreting the results of the evaluation and analysis to better understand the NLUs' confidence estimation and get more insights about our calibration measures.

6.1 Validity of Calibration Measures

In this study, we conduct an evaluation on model and rank levels by applying i) reliability diagrams and ii) correlation of confidence estimates with respect to instance-level accuracy. In order to assess the validity and relevance of the chosen evaluation measurements, we look at the extent to which results from the different measurements resonate with each other.

The model-level reliability diagram seems to resonate with the model-level Spearman's correlations. For instance, Rasa-Sklearn shows the best calibration in the reliability diagram as well as the strongest monotonicity across NLUs. In addition, Watson shows an obvious miscalibration in the reliability diagrams and weak monotonicity.

However, we note that the Spearman's correlation is lower on a rank level than on a model level. This can be explained by the fact that ranks extend across smaller ranges of confidence estimates, which increases the noise (See model-level Histogram 4b in comparison with rank-level Histograms 5b,6b,7b,8b). Generally, when using Spearman's correlation between confidence estimates and instance-level accuracy, a higher Spearman's correlation coefficient may be caused by stronger monotonicity, but can also be due to a larger variation in the confidence estimates, hence, less noise.

Additionally, a dissonance has been observed in rank-level calibration. While Spearman's correlation results suggest a decrease in the calibration with the descend of the rank, the rank-level reliability diagrams show that Rasa-Sklearn and Rasa-DIET have better calibration in lower ranks. This dissonance can be caused by the range factor that causes lower coefficients on rank level than model level, that is, lower ranks have smaller confidence ranges and less variance. Therefore, comparisons of the monotonicity and differences in Spearman's correlation may be hard to interpret across ranks, especially when the range and variation of confidence estimates vary in each rank.

6.2 Interpreting Calibration and Potential Applications

Can our model-level quantitative results be used to rate calibration in some absolute sense (e.g. coefficient X is good and Y is bad)? This seems difficult at this stage since no previous research – as far as we know – used Spearman's correlation to measure the correlation of confidence with respect to instance-level accuracy[9]. Nevertheless, they enable comparisons of degree of calibration between NLUs.

On a model level, monotonicity is viewed as a characteristic of well-calibrated NLUs. The stronger the monotonicity, the more reliable the ranking of hypotheses in a prediction. This enables, inter alia, thresholding of hypotheses when processing and applying semantic clarification on the output of the

[9]Dong et al. (2018) uses Spearman's correlation between confidence and instance-level F1 score rather than accuracy, and Vasudevan et al. (2019) uses Pearson's correlation rather than Spearman's correlation between confidence and instance-level accuracy.

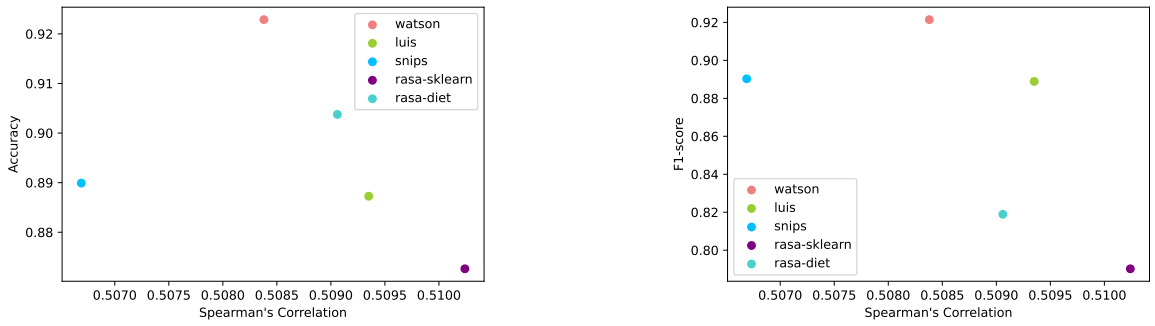
NLU.

Differences in the degree of calibration across ranks has been observed for all NLUs. Specifically, several of the NLUs are better calibrated for higher-ranked hypotheses than for lower-ranked ones. For dialogue systems developers, we may interpret this as indicating that it may be useful to look at the top two or three hypotheses when trying to detect ambiguity in input utterances. Looking at hypotheses ranked lower than 4 is likely to not be very informative. Fortunately, ambiguities are much more frequently 2-way (i.e. there are two possible interpretations of an input) or 3-way than 4-way or more. Knowledge about calibration is potentially useful for any downstream task that relies on confidence estimates associated with NLU hypotheses, such as choice of grounding strategies, ambiguity detection or re-scoring of hypotheses based on contextual information not available to the NLU but to the dialogue manager (such as dialogue state).

6.3 Performance vs. Calibration

When plotting the model-level performance and correlation, no clear pattern can be seen (see Figure 11). Our results suggest that the best calibrated NLU is Rasa-Sklearn and the poorest calibrated NLU is Snips, while Watson outperforms the NLUs and Rasa-Sklearn shows the worst performance.

A consequence of this is that when it comes to choosing an NLU for a dialogue system, there is likely to be a trade-off between performance (good for getting the right interpretation) and calibration (good for deciding on grounding stratigies and detecting input that is ambiguous from the NLU perspective). In fact, such trade-off between calibration and performance has been previously observed for neural networks (Guo et al., 2017). A potential choice between good performance or good calibration can depend on the domain and context of the dialogue system, where some domains may require better performance while other domains prioritize a reliable estimation of the hypotheses' likelihood.



(a) Accuracy vs. Spearman's Correlation

(b) F1-score vs. Spearman's Correlation

Figure 11: (Model-level) Accuracy (a) and F1-score (b) vs. Spearman's Correlation.

7 Ethics and Validity Threats

7.1 Construct Validity

Since there is no previous research – to the best of our knowledge – that has evaluated the calibration of NLUs, we faced a construct validity threat when planning the evaluation process and choosing the evaluation methods. To mitigate this threat, we extend established calibration measurement methods that have been successfully used in classification tasks, and adapt them to NLUs on model and rank levels. However, we observe some limitations of Spearman's correlation on a rank level, especially on lower

ranks where the correlation coefficent doesn't necessarily represent the calibration due to the presence of other factors like the range and variance of the confidence estimates (explained in Section 6.1). Therefore, further exploration is needed into how calibration of NLUs can be measured

7.2 Internal Validity

The selection of NLUs in this study introduced many parameters that could possibly change our results. The NLUs in general and especially Rasa can change their behaviour with the change of their configurations. Therefore, we use the default configurations for consistency across NLUs. We argue that our configuration choice is what an average dialogue developer would apply when using the evaluated NLUs.

Furthermore, None/irrelevant intents were removed from the output of NLUs, decreasing internal validity since the absence of the "None" hypothesis may have an impact on the model-level confidence estimates. Normalizing the confidence estimates in the intent ranking (after removing the None intent) may seem like a mitigation. However, this caused issues due to Watson being a multi-binary classifier, and we lack knowledge of how confidence estimates are calculated for each NLU, making the scores not suitable for normalization.

7.3 External Validity

Evaluating machine-learning based NLUs involves splitting data into training and test sets. This introduces a risk that obtained results depend on a specific split. In order to estimate and partly mitigate this risk, we perform repeated random sub-sampling as a Cross-validation method with 10 iterations (splits), we then average the results across iterations and provide the standard deviation. Our results show a low standard deviation across splits which increases the generalizability of our findings. Moreover, we enable reproducability and replicability of our evaluation by providing all the scripts, training and test sets and requirements.

7.4 Data Fallacies

Throughout our study, we acknowledge the importance of being transparent, fair and unbiased to make the evaluation trustworthy, and to avoid inflicting damage on the evaluated parties.

We attempt to minimize the sampling bias by i) considering a multi-domain dataset rather than a context-specific dataset, and ii) selecting a representative subset that maximizes the number of domains, scenarios, intents and examples covered. In addition, we mitigate the Danger of Summary Metrics by i) considering the raw results of all splits (Appendix 9.2, 9.3, 9.4) along with the standard deviations in our data analysis and discussion and ii) analysing the results on a rank level which has provided a more detailed view on the model-level evaluation.

Finally, we do not claim nor draw a final conclusion in RQ3 in regard to the correlation between calibration and performance due to the small sample size that didn't allow us to perform a statistical test. We only present a scatter plot that shows the absence of a clear trend or pattern which does not indicate a potential correlation between NLU calibration and performance.

8 Conclusion and Future Work

We took a methodology for evaluating the calibration of neural networks in intent classification tasks (Vasudevan et al., 2019; Guo et al., 2017) and applied it to NLU, in order to measure the calibration of 5 state-of-the-art NLUs, and evaluate their performances. We also extended the methodology to look at hypotheses on all ranks in the intent ranking on rank level (results per rank) and a model level (results of aggregated ranks).

Our findings show that on a model level, Rasa-Sklearn is the best calibrated NLU and Snips with the poorest calibration. On a rank level, the calibration decreases in lower ranks for Watson, LUIS and Snips, and vice versa for Rasa. We also highlight a trade-off between calibration and performance where Rasa-Sklearn – the best calibrated model – had the worst performance.

Future work can involve adapting our evaluation methods to detect ambiguity. For example, given an utterance that has two possible intents (hence ambiguous), a well-calibrated NLU should be able to assign similar confidence estimates to the two possible intents.
We also encourage further improvement of our rank-level quantitative analysis by applying another measure of calibration, e.g., Brier score (square loss), Log loss (Kull et al., 2017) or the improved ECE approach by Nixon et al. (2019). The results of the loss and/or the error of the NLU confidence estimates in regard to the instance-level accuracy would widen the scope of the results and may be more relevant for assessing rank-level calibration of the NLUs.

References

Abdellatif, A., Badran, K., Costa, D., & Shihab, E. (2021). A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*.

Ashukha, A., Lyzhov, A., Molchanov, D., & Vetrov, D. (2020). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*.

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Braun, D., Mendez, A. H., Matthes, F., & Langen, M. (2017). Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 174–185).

Bunk, T., Varshneya, D., Vlasov, V., & Nichol, A. (2020). Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*.

Canonico, M. & De Russis, L. (2018). A comparison and critique of natural language understanding tools. *Cloud Computing*, 2018, 120.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al. (2018). Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Dong, L., Quirk, C., & Lapata, M. (2018). Confidence modeling for neural semantic parsing. *arXiv preprint arXiv:1805.04604*.

Dubitzky, W., Granzow, M., & Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

Gregori, E. (2017). Evaluation of modern tools for an omscs advisor chatbot. *SMARTech: smartech.gatech.edu*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning* (pp. 1321–1330).: PMLR.

IBM (2010). Ibm watson. Online available at: `https://www.ibm.com/watson`. Accessed on: 2022-04-14.

Jung, S. (2019). Semantic vector learning for natural language understanding. *Computer Speech & Language*, 56, 130–145.

Kar, R. & Haldar, R. (2016). Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*.

Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S. (2018). Motivations, classification and model trial of conversational agents for insurance companies. *arXiv preprint arXiv:1812.07339*.

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning* (pp. 2796–2804).: PMLR.

Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics* (pp. 623–631).: PMLR.

Liu, X., Eshghi, A., Swietojanski, P., & Rieser, V. (2021). Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction* (pp. 165–183). Springer.

McTear, M., Callejas, Z., & Griol, D. (2016). The conversational interface: Talking to smart devices: Springer international publishing. *Doi: https://doi. org/10.1007/978-3-319-32967-3*.

Microsoft (2017). Luis (language understanding) - cognitive services. Online available at: `https://www.luis.ai/home`. Accessed on: 2022-04-14.

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. *CVPR Workshops*, 2(7).

Rasa (2016). Rasa: Open source conversational ai. Online available at: `https://rasa.com/`. Accessed on: 2022-04-14.

Runeson, P. & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2), 131–164.

Shridhar, K., Dash, A., Sahu, A., Pihlgren, G. G., Alonso, P., Pondenkandath, V., Kovács, G., Simistira, F., & Liwicki, M. (2019). Subword semantic hashing for intent classification on small datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6).: IEEE.

Snips (2013). Snips.ai. Online available at: `https://snips.ai/`. Accessed on: 2022-04-14.

Stoyanchev, S., Lison, P., & Bangalore, S. (2016). Rapid prototyping of form-driven dialogue systems using an open-source framework. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 216–219).

Tur, G. & De Mori, R. (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Vasudevan, V. T., Sethy, A., & Ghias, A. R. (2019). Towards better confidence estimation for neural models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7335–7339).: IEEE.

Wang, Y.-Y., Deng, L., & Acero, A. (2005). Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5), 16–31.

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14), 3866–3878.

9 Appendix

9.1 Rasa Pipelines

```
version: '2.0'

language: en

pipeline:
  - name: SpacyNLP
    model: "en_core_web_md"
    case_sensitive: False
  - name: SpacyTokenizer
    "intent_tokenization_flag": False
    "intent_split_symbol": "_"
    "token_pattern": None
  - name: SpacyFeaturizer
    "pooling": "mean"
  - name: "SklearnIntentClassifier"
    kernels: ["linear"]
    "gamma": [0.1]
    "max_cross_validation_folds": 1
    "scoring_function": "f1_weighted"
```

Listing 2: The configuration of Rasa-Sklearn pipeline used in this study.

```
version: '2.0'

language: en

pipeline:
  - name: WhitespaceTokenizer
  - name: RegexFeaturizer
  - name: LexicalSyntacticFeaturizer
  - name: CountVectorsFeaturizer
  - name: CountVectorsFeaturizer
    analyzer: char_wb
    min_ngram: 1
    max_ngram: 4
  - name: DIETClassifier
    epochs: 10
```

Listing 3: The configuration of Rasa-DIET pipeline used in this study.

9.2 Reliability Diagrams with Standard Deviation

We include the reliability diagrams we present in Section 5 with the standard deviation of mean accuracies for splits plotted. We provide the model-level reliability diagram in Figure 12 and rank-level reliability diagrams in Figure 13.

9.3 Rank-level Ranks vs Spearman's Correlations Plot with Standard Deviation

This Appendix includes the plot of the rank-level Spearman's correlation we presented in Section 5 in Figure 10 with the standard deviation of splits plotted in Figure 14.

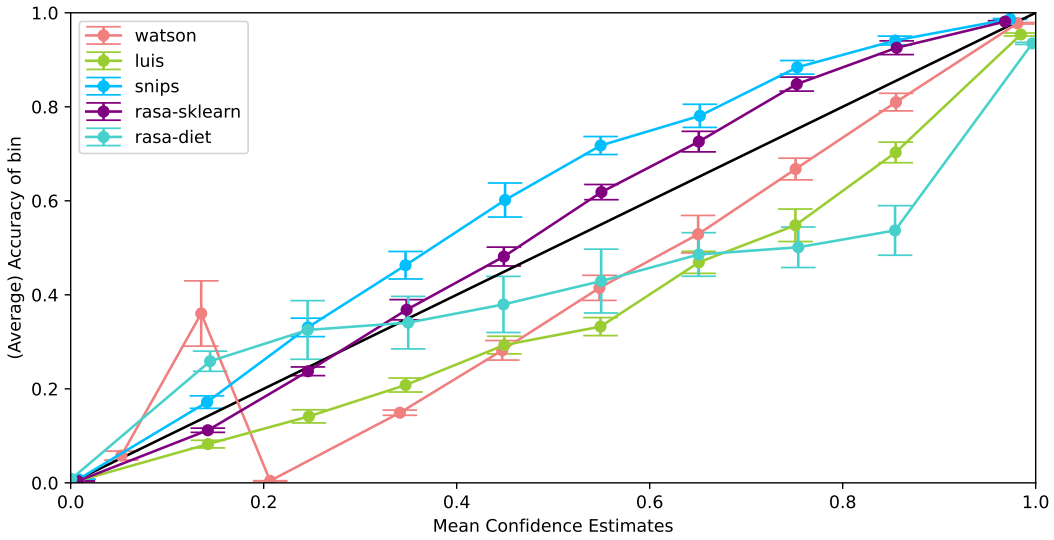


Figure 12: Model-level reliability diagram for Watson, LUIS, Snips, Rasa-Sklearn and Rasa-DIET with standard deviation.

9.4 Results of Rank-level Spearman's Correlation

In Section 5, we plot the rank-level Spearman's correlation of all ranks in Figure 10. In this section, we provide the raw data of the averaged Spearman's correlation across splits in Table 8 on rank 1-10 along with the standard deviation of splits.

NLU	**Rank**	**Spearman's Correlation (ρ)**	***p*-value**	**Standard Deviation**
Watson	1	0.36478	<0.0001	0.00588
LUIS	1	0.39553	<0.0001	0.01253
Snips	1	0.3908	<0.0001	0.00661
Rasa-Sklearn	1	0.41435	<0.0001	0.01068
Rasa-DIET	1	0.40187	<0.0001	0.00785
Watson	2	0.27888	<0.0001	0.01051
LUIS	2	0.29684	<0.0001	0.01383
Snips	2	0.25592	<0.0001	0.00891
Rasa-Sklearn	2	0.34423	<0.0001	0.01026
Rasa-DIET	2	0.31463	<0.0001	0.00987
Watson	3	0.15082	<0.0001	0.01208
LUIS	3	0.18616	<0.0001	0.00713
Snips	3	0.15376	<0.0001	0.01087
Rasa-Sklearn	3	0.19943	<0.0001	0.00566
Rasa-DIET	3	0.18482	<0.0001	0.00728
Watson	4	0.08652	<0.0001	0.00791
LUIS	4	0.12456	<0.0001	0.00807
Snips	4	0.10944	<0.0001	0.01031
Rasa-Sklearn	4	0.10235	<0.0001	0.01723
Rasa-DIET	4	0.11088	<0.0001	0.01128
Watson	5	0.05189	<0.0001	0.01222
LUIS	5	0.08779	<0.0001	0.01159

Snips	5	0.07635	<0.0001	0.0137
Rasa-Sklearn	5	0.06645	<0.0001	0.01872
Rasa-DIET	5	0.08253	<0.0001	0.01207
Watson	6	0.05453	<0.0001	0.01664
LUIS	6	0.07276	<0.0001	0.0116
Snips	6	0.04989	<0.0001	0.00622
Rasa-Sklearn	6	0.04359	<0.0001	0.01043
Rasa-DIET	6	0.06002	<0.0001	0.01147
Watson	7	0.05031	<0.0001	0.00933
LUIS	7	0.05709	<0.0001	0.00833
Snips	7	0.03705	<0.0001	0.01007
Rasa-Sklearn	7	0.04221	<0.0001	0.00658
Rasa-DIET	7	0.04454	<0.0001	0.01123
Watson	8	0.04464	<0.0001	0.01699
LUIS	8	0.03457	<0.0001	0.0088
Snips	8	0.02296	<0.0001	0.00871
Rasa-Sklearn	8	0.0272	<0.0001	0.01406
Rasa-DIET	8	0.03761	<0.0001	0.00858
Watson	9	0.0271	<0.0001	0.01933
LUIS	9	0.02164	<0.0001	0.00826
Snips	9	0.0277	<0.0001	0.00801
Rasa-Sklearn	9	0.02503	<0.0001	0.00858
Rasa-DIET	9	0.03419	<0.0001	0.0112
Watson	10	0.02725	<0.0001	0.01198
LUIS	10	0.01272	<0.0001	0.02247
Snips	10	0.01393	<0.0001	0.0141
Rasa-Sklearn	10	0.01372	<0.0001	0
Rasa-DIET	10	0.02131	<0.0001	0.0218

Table 8: Spearman's Correlation Coefficient: averaged rank-level NLU correlation for 10 splits with the *p*-value and the standard deviation of the splits correlation coefficients.

9.5 Post-hoc Analysis: t-test Calculations

In this Appendix, we present our statistical test results using t-test and Cohen's d on Spearman's Correlation results (Model level in Table 9 and rank level in Table 10), and performance results in Table 11.

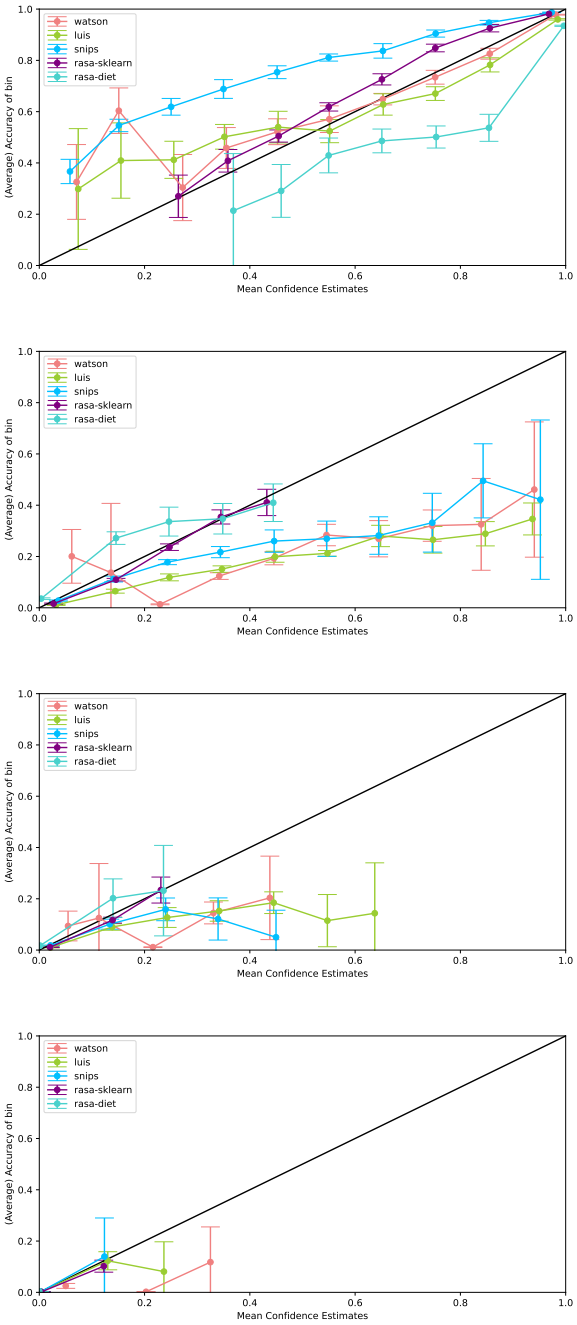


Figure 13: Rank-level reliability diagrams for ranks 1 (top), 2, 3 and 4-10 (bottom) with standard deviation.

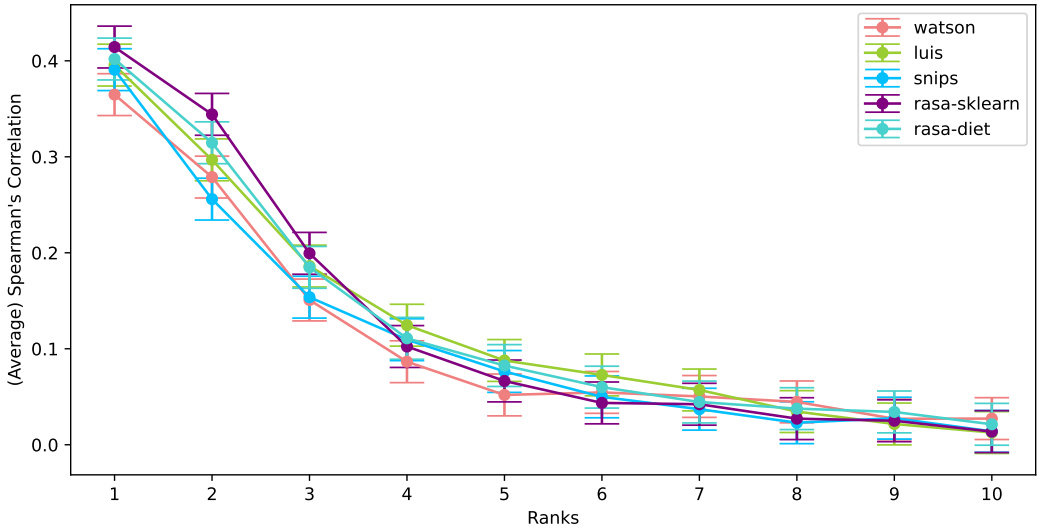


Figure 14: Spearman's Correlation for ranks 1-10 (with standard deviation).

Pairwise Comp.	**t-Statistic**	***p*-value**	**df**	**Effect Size**	**SSD (*p*<.05)**
(Watson, LUIS)	-3.1645	0.01147	9	L	Yes
(Watson, Snips)	4.9025	0.00084	9	L	Yes
(Watson, Rasa-Sklearn)	-5.4977	0.0003813	9	L	Yes
(Watson, Rasa-DIET)	-2.9555	0.01608	9	L	Yes
(LUIS, Snips)	25.569	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-3.8306	0.00402	9	L	Yes
(LUIS, Rasa-DIET)	-78.645	0.2895	9	S	No
(Snips, Rasa-Sklearn)	-16.545	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-7.8118	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-4.1319	0.002552	9	L	Yes

Table 9: t-test on pairwise NLUs' Spearman's correlation scores on a model level to determine if there is a statistical significant difference (SSD) and Cohen's d to measure the effect size (L- Large, M- Moderate, S- Small, N- Negligible).

Pairwise Comp.	**t Statistics**	***p*-value**	**df**	**Effect Size**	**SSD (*p*<.05)**
		Rank 1			
(Watson, LUIS)	-7.6715	<0.00001	9	L	Yes
(Watson, Snips)	-9.7613	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	-11.441	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-10.782	<0.00001	9	L	Yes
(LUIS, Snips)	1.2402	0.2463	9	S	No
(LUIS, Rasa-Sklearn)	-4.45	0.0016	9	L	Yes
(LUIS, Rasa-DIET)	-1.8668	0.09477	9	M	No
(Snips, Rasa-Sklearn)	-5.7598	0.0002729	9	L	Yes
(Snips, Rasa-DIET)	-3.0576	0.01362	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-6.754	<0.00001	9	L	Yes
		Rank 2			
(Watson, LUIS)	-3.2206	0.01048	9	L	Yes
(Watson, Snips)	6.4881	0.000113	9	L	Yes
(Watson, Rasa-Sklearn)	-17.398	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-8.6273	<0.00001	9	L	Yes
(LUIS, Snips)	9.9936	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-9.7455	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	-3.7508	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	-17.882	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-12.898	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-11.323	<0.00001	9	L	Yes
		Rank 3			
(Watson, LUIS)	-6.7607	<0.00001	9	L	Yes
(Watson, Snips)	-0.6851	0.5105	9	S	No
(Watson, Rasa-Sklearn)	-13.616	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	-6.2648	0.000147	9	L	Yes
(LUIS, Snips)	7.0407	<0.00001	9	L	Yes
(LUIS, Rasa-Sklearn)	-6.3356	0.0001352	9	L	Yes
(LUIS, Rasa-DIET)	0.46202	0.655	9	N	No
(Snips, Rasa-Sklearn)	-11.323	-7.0872	9	L	Yes
(Snips, Rasa-DIET)	-7.0872	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	-4.6652	0.001177	9	L	Yes
		Rank 4-10			
(Watson, LUIS)	-5.9362	<0.00001	49	L	Yes
(Watson, Snips)	-0.72951	0.4692	49	N	No
(Watson, Rasa-Sklearn)	0.078179	0.938	49	N	No
(Watson, Rasa-DIET)	-3.3111	0.00175	49	S	Yes
(LUIS, Snips)	9.1052	<0.00001	49	L	Yes
(LUIS, Rasa-Sklearn)	8.087	<0.00001	49	L	Yes
(LUIS, Rasa-DIET)	3.9641	0.0002393	49	M	Yes
(Snips, Rasa-Sklearn)	1.2524	0.2164	49	N	No
(Snips, Rasa-DIET)	-4.1725	0.0001228	49	M	Yes
(Rasa-DIET, Rasa-Sklearn)	5.2551	<0.00001	49	M	Yes

Table 10: t-test on pairwise NLUs' Spearman's correlation scores on a rank level to determine if there is a statistical significant difference (SSD) and Cohen's d to measure the effect size (L- Large, M- Moderate, S- Small, N- Negligible).

Pairwise Comp.	**t Statistics**	***p*-value**	**df**	**Effect Size**	**SSD (*p*<.05)**
		Accuracy			
(Watson, LUIS)	18.462	<0.00001	9	L	Yes
(Watson, Snips)	29.325	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	25.059	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	12.82	<0.00001	9	L	Yes
(LUIS, Snips)	-0.62904	0.545	9	N	No
(LUIS, Rasa-Sklearn)	11.672	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	-7.2468	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	13.889	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	-7.7684	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	18.968	<0.00001	9	L	Yes
		F1-score			
(Watson, LUIS)	15.437	<0.00001	9	L	Yes
(Watson, Snips)	25.432	<0.00001	9	L	Yes
(Watson, Rasa-Sklearn)	79.213	<0.00001	9	L	Yes
(Watson, Rasa-DIET)	73.47	<0.00001	9	L	Yes
(LUIS, Snips)	1.1095	0.296	9	S	No
(LUIS, Rasa-Sklearn)	95.383	<0.00001	9	L	Yes
(LUIS, Rasa-DIET)	49.549	<0.00001	9	L	Yes
(Snips, Rasa-Sklearn)	135.47	<0.00001	9	L	Yes
(Snips, Rasa-DIET)	88.435	<0.00001	9	L	Yes
(Rasa-DIET, Rasa-Sklearn)	18.098	<0.00001	9	L	Yes

Table 11: t-test on pairwise NLUs' Performance to determine if there is a statistical significant difference (SSD) and Cohen's d to measure the effect size (L- Large, M- Moderate, S- Small, N- Negligible).