DSAIT 4090 Group Project: Numerical & Temporal Fact Checking

Moniek Smink
Sowmya Prakash
Szymon Gołębiowski
Teodor Stanilla

1 Introduction

As the world continues to digitalize at a rapid pace, the use of digital tools to spread misinformation has also increased tremendously (Anderson and Rainie, 2017). Misinformation has serious consequences for society, such as inciting uncertainty, perpetuating violence, and initiating turmoil (Vosoughi et al., 2018). Manual fact-checking is a burdensome and practically unmanageable task in today's world. Thus, automatic fact-checking is being explored, particularly in the field of AI (Guo et al., 2022).

In this report, we will focus on numerical claims, as these claims are particularly persuasive according to the Illusion-of-Numerical-Truth effect and have not been the primary focus of much of the existing fact-checking research (Sagara, 2009; V et al., 2024). Inspired by (V et al., 2024), we define numerical claims as those which contain explicit or implicit numerical/temporal information and may require numerical operations for verification. We further break down numerical claims into several types including, statistical, temporal, interval, and comparison claims. To accomplish this, we will use the QuanTemp dataset, consisting of 7,302 statistical claims, 4,193 temporal claims, 2,357 interval claims, and 1,645 comparison claims.

The numerical fact-checking task usually consists of three stages: claim detection, evidence retrieval/ranking, and veracity prediction (V et al., 2024). A numerical claim is detected, relevant evidence is retrieved, and a veracity predictor predicts whether the evidence supports, refutes, or conflicts with the evidence (Guo et al., 2022). We will focus on the veracity prediction portion of numerical fact-checking. Veracity prediction is a natural language inference (NLI) task which takes as input a claim and a set of evidence and predicts a label: True, False, or Conflicting.

The veracity prediction task's structured nature allows most state-of-the-art language models to be used as the predictor's backbone. While general state-of-the-art language models have shown promise in numeric fact-checking, there are specific numerically inclined language models that also show promise (V et al., 2024). We define a numerically inclined language model as a language model that has been specifically adjusted to perform numerical tasks.

Another enhancement to fact-checking is when a claim is broken down or decomposed into a set of sub-questions. This is called claim decomposition. Natural language claims are often complex, particularly when they are numerical in nature. Thus, significant improvements in fact check performance are seen when first decomposing these claims, retrieving evidence relevant to each sub-question, then performing veracity prediction on these decomposed claims and associated evidence (V et al., 2024).

We aim to evaluate current general & numerically inclined language models on the veracity prediction of different types of numerical claims in a quantitative and qualitative analysis. Furthermore, we also aim to evaluate and explore claim decomposition within the numerical fact-checking task. We accomplish these aims by answering the following research questions.

RQ1: How do existing state-of-the-art NLI models including LLM based models perform on different types of numerical claims in QuanTemp?

RQ2: Does special handling of numerical information improve performance for NLI models (fine-tuned and LLM based)?

RQ3: Does claim decomposition help in better evidence retrieval and downstream veracity prediction?

2 Methods[1]

2.1 Dataset

The QuanTemp dataset (V et al., 2024) is used to carry out experiments on various NLI models for Numerical Fact Checking. The dataset is comprised of 15,514 numerical claims sourced from various fact checking platforms like Politifact, Snopes, AfP, etc. The dataset also contains an evidence corpus of 423,320 snippets, which can be used to verify claims. In addition, each QuanTemp claim has some 'gold standard evidence' that comes from the fact-checking websites. The claim types are categorized into Statistical, Temporal, Interval and Comparison claims. The dataset is split into train, validation and test sets with 9,935, 3,084 and 2,495 claims in each set respectively. The label distribution for true claims is 18.79%, false claims is 57.93% and conflicting claims is 23.27%.

During this project, we found several curiosities within the dataset. First, we noticed several claims within the different categories that did not fit the definition of those categories. For example, the claim: "The video shows G20 delegates having lunch with their hands" was labeled as temporal. Furthermore, we found several claims that read more as questions, such as "Does This Video Show Moldova Residents Blocking A NATO Column In 2022?". An interesting thing to note is that, although the QuanTemp paper claims to have filtered out non-English claims, we found several claims written in non-English languages (mainly German). For comparability with the QuanTemp paper, we decided not to filter out any non-English or curious claims we found.

2.2 Evidence Retrieval

For testing our models for RQ1 and RQ2, to compare model performance in a frozen evidence retrieval state, the gold standard evidence is used for each QuanTemp claim. We chose to use the gold standard evidence because it contains complex, coherent pieces of text that require numeric reasoning to decipher. An example excerpt of a random piece of gold standard evidence for the claim "Every taxpayer owes about $130,000 to pay off the national debt." can be read here:

> "[...] estimated the debt at close to $14.3 trillion and had the debt per taxpayer at $128,371. It says there are 111,087,453 U.S. taxpayers with taxable income. The people who maintain the website apparently divided the debt total by that 111 million figure to come up with that $128,371 total. We tried to ask them how specifically they came up with that total. The website had no telephone number. We sent them an e-mail and got an automatic response [...]"

For RQ3, we leverage a multi-step evidence retrieval process on different stages of claim decomposition. First, we retrieve the top-100 documents from the corpus using BM25 (Robertson and Zaragoza, 2009) on the original QuanTemp claim. Then, we leverage a paraphrase-MiniLM-L6-v2 ranking model from the sentence-transformers library (Reimers and Gurevych, 2019) on either the original claim or the decomposed sub-questions generated during claim decomposition.

For evidence retrieval on the original claim, we return the top-k results where k is within 0 to 3 depending on whether the ranking scores are above an arbitrary threshold of 0.5 to prevent irrelevant evidence from confusing the model. For evidence retrieval on decomposed sub-questions, we take the top-1 results for each sub-question, remove duplicates and any result where the ranking score is below the arbitrary threshold. If the number of evidences after this removal is less than 3, we repeat this process with the top-2 results. Then with the top-3 results, meaning that for evidence retrieval on decomposed sub-questions a claim may be supported by 0-5 evidences depending on the duplication and scores of the top-3 evidences for each sub-question. This process allows the most relevant evidences for each sub-question to appear, preventing overcrowding by one sub-question, but still allowing flexible evidencing if one sub-question has no relevant evidences at all.

2.3 Veracity Prediction Models

2.3.1 General Language Models

To examine veracity prediction for more general language models, we chose a selection of pretrained natural language inference (NLI) models including Roberta-Large-MNLI (Liu et al., 2019), Bart-Large-MNLI (Lewis et al., 2019), and Deberta-v3-Base-Tasksource-NLI (He et al., 2023),

[1]Code for our report can be found at: https://github.com/sgol13/tudelft-nlp-fact-checking; Data is stored in the Google Drive directory: https://drive.google.com/drive/folders/1QOwZL0Kaov1enVqq2JTpuwDGuuTTfs-5?usp=sharing

as well as some generative language models including Flan-T5-Base (Chung et al., 2022), and GPT2 (Radford et al., 2019).

The RoBERTa-Large-MNLI and Bart-Large-MNLI models are RoBERTa-Large (Liu et al., 2019) and Bart-Large (Lewis et al., 2019) models fine-tuned on the Multi-Genre Natural Language Inference corpus (Williams et al., 2018). The DeBERTa-v3-Base-Tasksource-NLI model is a DeBERTa-v3-base model (He et al., 2021) fine-tuned with multi-task learning on 600+ tasks in the tasksource collection (Sileo, 2024). The Flan-T5-Base model is an enhanced version of T5-Base (Raffel et al., 2023) that has been fine-tuned in a mixture of tasks. The GPT2 model is a next-word generative model from OpenAI.

Each model was fine-tuned on the QuanTemp dataset using gold standard evidence with an AdamW optimizer and a learning rate of 2e-5 till 'EarlyStopping' with a patience of 2. Batchsizes varied between 4 to 16 depending on the computational resources available. The first five layers of each model were frozen during fine-tuning for time complexity and prevention of model forgetting. Training was mostly done using Google T4, L4, or A100 GPUs or locally with various GPUs as available.

2.3.2 Numerically Inclined Language Models

To evaluate whether more numerically inclined language models could perform better in numerical fact checking, we chose a selection of pre-trained models including MathRoBERTa, Numeric-T5 (Yang et al., 2021), PASTA (Gu et al., 2022), and Elastic RoBERTa (Zhang and Moshfeghi, 2022).

MathRoBERTa is a fine-tuned RoBERTa-Large model (Liu et al., 2019) on 3 million math discussion posts. Numeric-T5 is a T5-Small model (Raffel et al., 2023) fine-tuned on four datasets designed to strengthen skills necessary for numerical reasoning over text and general reading comprehension including the Discrete Reasoning over Text (DROP) dataset (Dua et al., 2019). PASTA performs table-based fact verification by pre-training a DeBERTa-V3 model (He et al., 2021) with different types of synthesized sentence-table cloze questions. Elastic RoBERTa has a RoBERTa (Liu et al., 2019) encoder and a four-module compiler which separates the handling of operators and operands allowing for more robust reasoning with diverse operators.

Like the general models, each numerically inclined model was fine-tuned on the QuanTemp dataset using gold standard evidence with an AdamW optimizer and a learning rate of 2e-5 till 'EarlyStopping' with a patience of 2. Batchsizes varied between 4 to 16 depending on the computational resources available. The first five layers of each model were frozen during fine-tuning for time complexity and prevention of model forgetting. Training was mostly done using Google T4 and L4 GPUs, or locally with various GPUs as available.

2.4 Claim Decomposition

Fact-checking claims are often complex with many moving parts. Claim decomposition, or the creation of sub-questions relating to an original claim, can aid a model during veracity prediction. For RQ3, we evaluate different claim decomposition methods. We used our veracity prediction model from each of our previous two research questions to serve as our test models for our claim decomposition methods.

As a baseline to compare our decomposition veracity performance to, we fine-tuned each test model with the gold standard evidence detailed in Section 2.1 and evidence retrieved based on the original claim following the procedure detailed in Section 2.2. We detail each of our claim decomposition methods in their associated subsections below. After obtaining the decomposed sub-questions, we fine-tune our test models with evidence retrieved from the sub-questions following the procedure detailed in Section 2.2.

2.4.1 CLAIMDECOMP

Our first claim decomposition method is derived from the CLAIMDECOMP dataset (Chen et al., 2022). This dataset is comprised of 1,000 complex original claims from PolitiFact. Each claim is accompanied by decomposed sub-questions written by fact checkers. We fine-tune a Flan-T5-Base (Chung et al., 2022) text generation model on the CLAIMDECOMP dataset to generate potential sub-questions for a claim. The model is fine-tuned for 36 epochs using a batch size of 8, max input token length of 64, max output token length of 128, learning rate of 2e-5, and a linear step size scheduler.

2.4.2 Pseudo Program-FC: In-Context GPT-3.5-Turbo

Our second claim decomposition method is inspired from Program-FC (Pan et al., 2023). Program-FC (Program-Guided Fact-Checking)

Mean	Std	Minimum	1st Quartile	Median	3rd Quartile	Maximum
31.38	34.33	0.00	0.00	12.50	62.50	100.00

Table 1: Summary statistics of the percentage of models fine-tuned on gold standard evidence that got a random claim wrong.

functions similarly to a piece code. It decomposes complex claims into simpler subtasks that are solved using specialized subtask handlers. At the root of Program-FC is its use of in-context learning, using OpenAI models such as GPT-3.5, to perform claim decomposition. Because our team did not have access to an OpenAI API key, we did not directly use OpenAI models to decompose our claims. Instead, we used the decompositions released with the QuanTemp dataset. Each original claim in the QuanTemp dataset was decomposed using GPT-3.5-Turbo. We use these decomposed claims as an approximation of the claims generated by Program-FC.

2.4.3 Custom Decomposition Method

In addition to the two claim decomposition methods detailed above, we wanted to implement a custom decomposition method, more specific to the types of claims found in the QuanTemp dataset. We fine-tuned a Bart-Large-MNLI model to predict the type of claim: statistical, temporal, interval, or comparison, as defined by the QuanTemp dataset. Then, we used a structured information extraction model called NuExtract-tiny-v1.5 (Cripwell, 2024) to extract quantitative information such as statistics and dates from the claim. NuExtract-tiny-v1.5 is a Qwen2.5-0.5B model (Yang et al., 2024) fine-tuned on a private synthetic structured information extraction dataset. Finally, in addition to the sub-questions generated by the CLAIMDECOMP method, we added custom sub-questions prompting the veracity prediction model whether the extracted numbers from NuRstract-tiny-v1.5 are rigorously supported by the evidence.

2.5 Qualitative Analysis: Random Claim Sampling

To evaluate the veracity prediction of our models we wanted to randomly sample some claims from the test set from a range of difficulties. To do this, we defined a claim's trickiness factor as the percentage of models fine-tuned on gold standard evidence that got the claim wrong. The summary statistics for this trickiness factor can be seen in Table 1. We then randomly sampled for each quartile of the trickiness factor a random claim each model got wrong. Please note that since the minimum and 1st quartile percentage are the same, we only sampled three total wrong claims for each model, not four.

3 Results

3.1 General LM Performance on Veracity Prediction

3.1.1 Quantitative Analysis

By comparing the macro and weighted F1 scores for each of our models, we see that, overall, the DeBERTa-v3-Base-Tasksource-NLI model performed the best on veracity prediction on the QuanTemp dataset (Table 2). We also see that the BART-Large-MNLI model performed better than the DeBERTa-v3-Base-Tasksource-NLI model for true claims and temporal claims. We see that across all models, temporal claims appear to be the most difficult type of claim to fact-check, while comparison claims are the easiest.

A Note About GPT-2: It must be mentioned that our GPT-2 fine-tuning was lackluster. No matter how we manipulated the hyperparameters or initialization, GPT-2 always fell into a local minima of predicting everything as false and never achieved comparable results to the other models. We tried fine-tuning the model on a much smaller subset of the data which it memorized perfectly, so we do believe our algorithm was learning something. However, whenever fine-tuning in environments similar to the other models, GPT-2 was not able to predict anything other than false for any QuanTemp claim. For this reason, we omit GPT-2 from the following qualitative analysis and claim decomposition results.

3.1.2 Qualitative Analysis

In addition to general metrics, we qualitatively examine the overall statistics and some specific false predictions for each model. In Figure 1 and 2 we see that different general models have different areas that need improvement. For RoBERTa-Large-MNLI, we see that it tends to label too many conflicting samples as false and has the trouble with

Model	Statistical		Temporal		Interval		Comparison		Per-class F1			QuanTemp	
	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	T-F1	F-F1	C-F1	M-F1	W-F1
RoBERTa-Large-MNLI	54.08	60.46	63.39	79.40	55.44	67.73	53.57	53.69	47.11	82.17	42.34	57.21	65.96
BART-Large-MNLI	59.66	64.58	**69.35**	**82.76**	58.3	68.72	56.24	54.58	**63.41**	84.54	37.00	61.65	69.13
DeBERTa-v3-Base-Tasksource-NLI	**63.83**	**68.15**	67.17	80.99	**63.28**	**73.51**	**60.97**	**60.16**	62.91	**84.81**	**47.33**	**65.02**	**71.67**
Flan-T5-Base	58.88	63.13	68.84	81.12	59.19	67.39	57.41	56.70	58.56	80.69	45.57	61.61	68.07
GPT-2 (Naive Majority Class)	22.47	34.26	28.31	62.66	25.87	49.2	16.52	16.32	0	72.64	0	24.21	41.43
MathRoBERTa	59.08	64.28	66.31	79.91	58.18	68.18	**60.76**	60.30	56.13	82.58	46.15	61.62	62.82
Numeric-T5	55.81	64.46	55.00	75.29	53.46	65.11	51.01	50.94	45.60	80.45	41.37	55.81	64.46
PASTA	51.16	57.24	42.39	69.50	46.07	61.13	52.16	50.76	48.20	80.81	23.68	50.90	60.92
Elastic RoBERTa	**63.65**	**68.04**	**72.16**	**83.00**	**64.94**	**72.19**	60.05	**60.55**	**58.72**	**83.52**	**60.87**	**65.78**	**72.00**

Table 2: Results fine-tuning and evaluating different general language models (top) and numerically inclined language models (bottom) on the QuanTemp dataset using gold standard evidence. Best results for each category and type of model are bolded.

mislabeling true samples in general. We see that BART-Large-MNLI has trouble with conflicting claims, often incorrectly labeling them as true. For DeBERTa-v3-Tasksource-NLI, we see that the predicted label distribution matches the ground truth distribution better than most other models, but the model still has trouble with the conflicting class. For Flan-T5 we see that it tends to be overconfident and incorrectly predict false claims as conflicting and conflicting claims as true. The trouble most models have with the conflicting class is likely due to its ambiguous and rare nature. It is likely a class that these models are not quite as familiar with when compared to the true and false classes.

These results are reinforced when we more closely examine some incorrect predictions for each model. Incorrectly predicted claims were randomly sampled from each quartile of the trickiness factor as described in Section 2.5. The results of this sampling can be seen in Table 3. We see that RoBERTa-Large-MNLI still has troubles with discerning true claims, even with gold standard evidence, even on claims in the bottom half of the trickiness factor. We see that BART-Large-MNLI, for this selection of random samples, tends to veer away from the hard true/false labels with gold standard evidence and goes for the safer alternative of conflicting. DeBERTa-v3-Tasksource-NLI made varied mistakes in this batch of random claims with gold evidence. We also see that Flan-T5-Base continues its trend to be overconfident in claims being true, even in claims that are below average for trickiness. This could be due to Flan-T5-Base having a max input token length of 256; perhaps the long gold standard is being cut-off before truly relevant evidence can be given.

3.2 Numerically Inclined LM Performance on Veracity Prediction

3.2.1 Quantitative Analysis

By comparing the macro and weighted F1 scores for each of our numerically inclined models, we see that, overall, the Elastic-RoBERTa model performed the best on veracity prediction on the QuanTemp dataset using gold standard evidence, significantly outperforming almost every other numerically inclined model (Table 2) and also slightly outperforming the best general language model (DeBERTa-v3-Base-Tasksource-NLI). Of the other numerically inclined models we see that Numeric-T5 and MathRoBERTa are not competitive with Elastic-RoBERTa, although still better than PASTA which has significantly lower performance.

3.2.2 Qualitative Analysis

In addition to general metrics, we qualitatively examine the overall statistics and some specific false predictions for each numerically inclined model. In Figure 1 and 2 we see that the different models have different areas that need improvement. For MathRoBERTa, we see that it tends to be too confident in incorrectly labeling false claims as conflicting or true. For Elastic-RoBERTa, we see that its predicted label distributions matches the ground truth distribution quite well, displaying the best performance in predicting conflicting labels as conflicting. However, Elastic-RoBERTa still has some trouble with incorrectly predicting true claims as conflicting. For PASTA, we see that it has an incredibly difficult time labeling any claim as conflicting. For Numeric-T5, we see that its label distribution quite closely matches the ground truth distribution, although it tends to predict too many true claims as

conflicting.

These results are perpetuated when we more closely examine some incorrect predictions for each model. Incorrectly predicted claims were randomly sampled from each quartile of the trickiness factor as described in Section 2.5. The results of this sampling can be seen in Table 3. We see that the MathRoBERTa model continues to have difficulty with false claims. We see that PASTA has some quite severe errors, mistaking true claims as false, and false claims as true. We see that Elastic-RoBERTa's errors tend to be closer in range (mistaking true as conflicting and conflicting as false) and quite understandable, especially when the random claim in the highest trickiness factor was completely in German. For Numeric-T5, we see an example of a true claim being predicted as conflicting.

3.2.3 Comparison to General LMs

In Table 2, we see that numerically inclined models do not significantly outperform general models, although our best model, Elastic-RoBERTa, a numerically inclined model, does outperform all others. This result is rather surprising because the QuanTemp paper claims that numerically inclined models do tend to outperform standard models (V et al., 2024). Perhaps this discrepancy can be explained by the fact that the general LMs evaluated in the QuanTemp paper were often prompted in-context (Flan-T5) or were the small versions (T5-small vs Numeric-T5-small). It must also be mentioned that the difference between BART-Large-MNLI and Elastic-RoBERTa in the QuanTemp paper was also very comparable to our result.

3.3 Claim Decomposition

3.3.1 CLAIMDECOMP Generative Claim Decomposition Model

Evaluation of the generative Flan-T5-Base claim decomposition model fine-tuned on the CLAIMDECOMP dataset can be seen in Table 4. Please note that although we do have decompositions of the QuanTemp claims from Pseudo Program-FC we do not have ground truth data to compare it to. Thus, this method is omitted from the quantitative results table (Table 4). Additionally, since the subquestions we add for our custom decomposition method are either generated by the CLAIMDECOMP method or statically added according to a classifier, this method is also omitted form the quantitative results table (Table 4).

Training curves for creating the claim decomposition generative Flan-T5-Base model fine-tuned on the CLAIMDECOMP dataset can be seen in Figure 3. We see that the ROUGE scores for the model improved with fine-tuning and had a reasonable peak around the 12th epoch of 36 total epochs.

An example of claim decompositions for each decomposition method can be seen in Table 5. We see that decompositions from CLAIMDECOMP tend to have more repetition and be more sporadic. Decompositions from Pseudo Program-FC are more focused and, when the claim is simple, are just a paraphrasing of the original claim. Finally, decompositions from the custom method have the same characteristics of the CLAIMDECOMP method along with one extra subquestion with the extracted numerical quantities.

3.3.2 Custom Decomp: Claim Classification Results

The classification accuracies from our BART-Large-MNlI model fine-tuned to classify a claim as statistical, temporal, interval, or comparison for the train, validation, and test sets within QuanTemp were 97.54%, 85.86%, and 85.81% respectively.

3.3.3 Veracity Prediction Results with Claim Decomposition

Quantitative Analysis: We compare the macro and weighted F1 scores for the test veracity prediction models, fine-tuned on subquestions and evidence retrieved from different decomposition methods. As expected, we see that our baseline gold standard evidence models perform the best (Table 6), with one exception. The gold standard evidence clearly and coherently states reasoning and supporting facts to verify the claim label; therefore, it follows that gold standard fine-tuned models would have the best performance in claim veracity prediction.

Of our decomposition methods, we see that, for three of the four general models (BART-Large-MNLI, DeBERTa-v3-Base-Tasksource-NLI, and Flan-T5-Base) and three of the four numerically inclined models (Numeric-T5, PASTA, & Math-RoBERTa), some kind of decomposition does improve the performance of the model over just retrieving evidence for the original claim alone. However, for one of the four general models (RoBERTa-Large-MNLI) and one of the four numerically inclined models (Elastic-RoBERTa), decomposition did not significantly help with veracity prediction.

Specifically, for BART-Large-MNLI, all three of the decomposition methods showed improvement over the original claim, with the CLAIMDECOMP method being the best, although to a slightly lesser extent than seen in the QuanTemp paper (V et al., 2024), likely because the sub-questions seen in the QuanTemp paper were generated using in-context learning on much more advanced LLMs (V et al., 2024). On the other hand, for RoBERTa-Large-MNLI, none of the decomposition methods came close to the original claim performance, although the custom method was the closest.

For DeBERTa-v3-Base-Tasksource-NLI, decomposing the claims using a Flan-T5-Base model fine-tuned on the CLAIMDECOMP dataset outperforms the original claim. However, the custom decomposition method actually had lower performance than the original claim which is curious considering that the custom method built off of the CLAIMDECOMP method. On the other hand, for the Flan-T5-Base model, the custom decomposition method actually improved the performance compared to baseline claim.

Math-RoBERTa showed the best performance with the custom decomposition method, showing that numerically inclined models can likely work better with our numerical in nature added subquestion. On the other hand, Numeric-T5 preferred the original CLAIMDECOMP method over the custom method, although all three decompositions performed higher than the original claim.

PASTA was the only model we explored that had better performance with non-gold standard evidence. This is likely due to the difficulty we had with training PASTA, as it was one of the largest models we trained and often failed its training. PASTA showed the best performance with the Pseudo Program-FC decomposition method, the only model to do so. PASTA's results with the CLAIMDECOMP and custom decomposition methods were lackluster. In general, we take these PASTA results with a grain of salt, as the training and results were inconsistent.

For Elastic-RoBERTa, our best veracity prediction model from RQ1 and RQ2, claim decomposition did not particularly help with the results, as none of the decomposition methods outperformed the original claim performance, although the custom method was the closest.

In general, we see that models who are overconfident (tend to predict more claims as true than the ground truth) benefit from the harsh subquestions found in our custom decomposition method, and models who are underconfident (tend to predict more claims as false than the ground truth) suffer from our custom decomposition method. Additionally, we see that numerically inclined models tend to have a better reaction to our custom decomposition method, likely because the added subquestion highlights the numerical nature of the claim.

Clearly, the different decomposition methods have different impacts on different models depending on their numerical understanding and their tendency towards overconfidence. Additionally, for nearly every model our CLAIMDECOMP or custom decomposition method outperformed the Pseudo Program-FC method, demonstrating that in-context learning may not always be the answer. However, we see that for our best veracity prediction model, Elastic-RoBERTa, claim decomposition did not lead to better veracity prediction.

Qualitative Analysis: In addition to the general quantitative metrics, we qualitatively examine the overall statistics for each test model fine-tuned with a different decomposition method. In Figure 4 and 5, we see that the different test models react differently to the different decomposition methods.

The RoBERTa-Large-MNLI model had the harshest reaction to the decomposition methods overall, switching from over-predicting claims as conflicting, to over-predicting claims as true, to being close to the original claim results with the custom decomposition method.

On the other hand, for BART-Large-MNLI, claim decomposition helped to stabilize the predictions for false and conflicting claims. Compared to the BART-Large-MNLI model fine-tuned using the gold standard evidence, the same model fine-tuned using evidence retrieved from the original claim shows a tendency to misclassify false claims and label conflicting claims as true. The BART-Large-MNLI model fine-tuned on sub-questions and evidences generated from the CLAIMDECOMP Flan-T5-Base model performs better than the original claim model but still shows tendencies to misclassify false claims more than the gold standard baseline. We see that the BART-Large-MNLI model with the Pseudo Program-FC tends to be overconfident in predicting all types of claims as false, specifically having trouble with labeling a true claim as true.

DeBERTa-v3-Base-Tasksource-NLI correctly la-

beled more conflicting claims with the CLAIMDECOMP method, correctly labeled more true claims with the Pseudo Program-FC and custom decomposition methods.

Flan-T5-Base showed almost no improvement with decomposition using CLAIMDECOMP or Pseudo Program-FC, always displaying difficulty with predicting any claim as true. However, there was a large jump in predicting true claims as true with the custom decomposition method.

For Math-RoBERTa, Pseudo Program-FC and custom decomposition caused more true claims to be predicted as true when compared to the original evidence version. Math-RoBERTa also displayed a resilience with and without claim decomposition, keeping a similar label balance among all decomposition methods.

On the other hand, claim decomposition greatly improved the label balance of Numeric-T5, correctly labeling more true claims as true, but also incorrectly labeling more conflicting claims as true.

As for PASTA, we can more clearly see that the training of this large model was irregular. For the gold standard evidence, the model rarely predicted any claim as false. For the CLAIMDECOMP and custom decomposition models we see that the model rarely predicted anything as true. The Pseudo Program-FC method appeared to create a happy middle between these extremes. In future work, the PASTA model should be examined more carefully to determine the root of these irregularities.

Our best model from RQ1 and RQ2, Elastic-RoBERTa, displayed almost no improvements with our decomposition methods, often being underconfident by predicting conflicting claims as false. Compared to the Elastic-RoBERTa model fine-tuned using gold standard evidence, the same model fine-tuned using original claim evidence shows a tendency to mislabel conflicting claims as true, although it does tend to correctly predict true claims as true. The Elastic-RoBERTa model fine-tuned on subquestions and evidences from the CLAIMDECOMP decomposition method has a similar problem, where it correctly predicts more true claims as true, but also incorrectly predicts more conflicting claims as true. For our custom decomposition method, Elastic-RoBERTa's underconfidence mixed harshly with the strict numerical subquestions, causing many more true and conflicting claims to be labeled as false.

4 Conclusions

We evaluate many language models in their ability to perform numerical fact checking on the QuanTemp dataset. We show that naive numerically inclined models are not always better than general knowledge language models in numerical veracity prediction. We show that claim decomposition during evidence retrieval does improve the performance of veracity prediction when retrieving non-gold-standard evidence for some models, but not all. We show that claim decomposition methods that are numerical in nature to tend to be better received by numerically inclined models. Finally, we show that QuanTemp is a challenging, though slightly flawed, dataset capable of challenging current state-of-the-art NLI models. If we had more time, our team would aim to spend more time refining our claim decomposition methods to work more universally across different models and we would want to explore more numerically inclined models such as LUNA.

5 References

References

Janna Anderson and Lee Rainie. 2017. The future of truth and misinformation online. *Pew Research Center*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. *arxiv preprint*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Liam Cripwell. 2024. Nuextract 1.5 - multilingual, infinite context, still small, and better than gpt-4o!

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-operations aware fact verification via sentence-table

cloze pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4971–4983, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Namika Sagara. 2009. Consumer understanding and use of numeric information in product claims.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Peng-Jian Yang, Ying Ting Chen, Yuechan Chen, and Daniel Cer. 2021. Nt5?! training t5 to perform numerical reasoning.

Jiaxin Zhang and Yashar Moshfeghi. 2022. Elastic: Numerical reasoning with adaptive symbolic compiler.

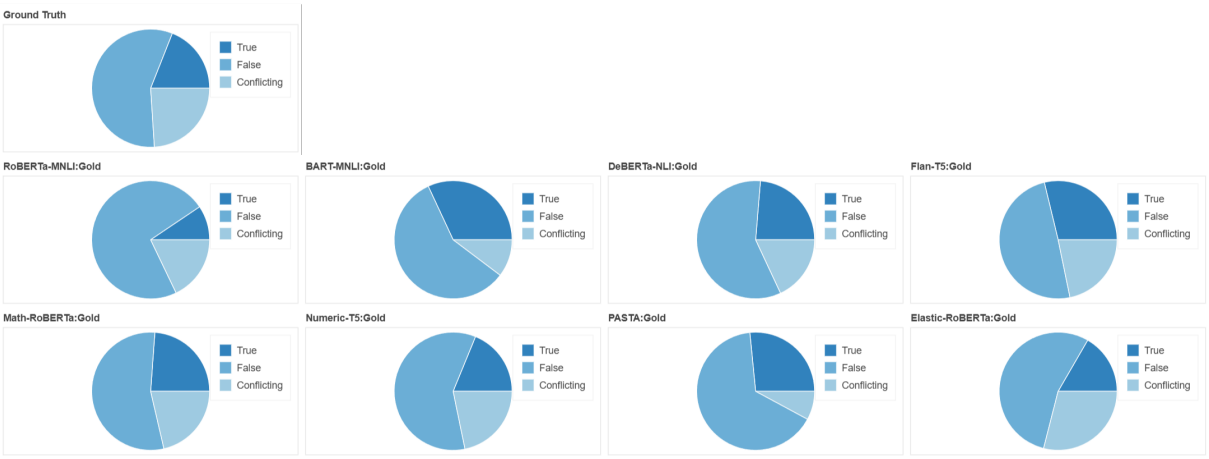

Figure 1: Pie charts of the distribution of predicted labels for each general and numerically inclined LM evaluated on the QuanTemp test set using gold standard evidence. Ground truth pie chart also included for comparison.

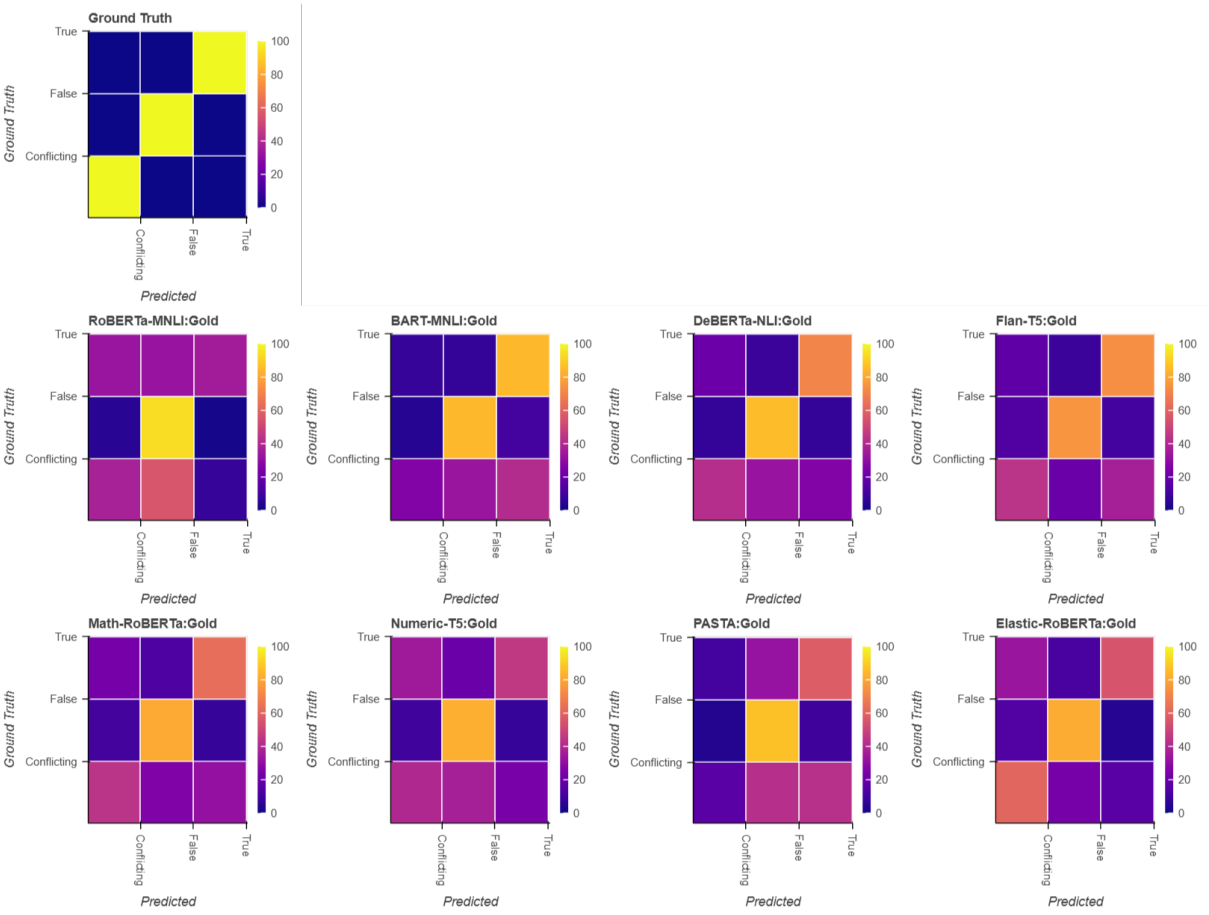

Figure 2: Confusion matrices for each general and numerically inclined LM evaluated on the QuanTemp test set using gold standard evidence. Ground truth confusion matrix also included for comparison. Values in the confusion matrix go from 0 to 100%, signifying that 100% of the samples truly in that class were predicted as such.

Model	Claim	Actual	Predicted
RoBERTa-Large-MNLI	"Every taxpayer owes about $130,000 to pay off the national debt."	True	Conflicting
	"My opponent, Rick Gunn, blocked the expansion of Medicaid - costing half a million people health insurance, including at least 23,000 veterans."	True	Conflicting
	Says Democratic recall opponent Mahlon Mitchell sent a letter "encouraging folks to boycott more than 100 companies."	Conflicting	False
BART-Large-MNLI	In New York, when she ran for reelection, she carried 58 of our 62 counties. George Bush had won 40 counties in New York just two years earlier.	True	Conflicting
	Some reports suggest Covid-19 vaccines have caused a 2,000% increase in miscarriages.	False	Conflicting
	"I think if you look to Chicago, where you had over 4,000 victims of gun-related crimes last year, they have the strictest gun laws in the country. That certainly hasn't helped there."	False	Conflicting
DeBERTa-v3-Tasksource-NLI	"Texas families have kept more than $10 billion in their family budgets since we successfully fought to restore Texas' sales tax deduction a decade ago."	True	Conflicting
	Says six studies verify that the math adds up for Mitt Romney's tax plan.	Conflicting	False
	"[The ANC has] increased the total values of goods and services produced in South Africa (GDP) 10-fold, over the 25-year period, from 1994 to 2019."	False	True
Flan-T5-Base	Coconut coir-infused water can treat typhoid in two weeks.	False	True
	"Gangs have increased by 40 percent since this president was elected."	Conflicting	True
	"Last financial year the Kenya Revenue Authority reported that the PAYE (Pay As You Earn) was KSh461 billion".	True	False
MathRoBERTa	"If you look at the three people on the (debate) stage from the United States Senate, all three of them have a combined two bills that became law that they've sponsored."	True	Conflicting
	"Today, we have two Vietnams, side by side, North and South, exchanging and working."	False	True
	Says "of the 2,000 Portland households in the year-long (composting) pilot, 87 percent of participants reported being satisfied with the overall system."	False	True
Numeric-T5	Wearing two Pnp face masks offers stronger protection than wearing one N95 mask.	False	True
	Says a portfolio managed by the Texas General Land Office earned 22 percent last year while the state's emergency reserve account experienced a 1 percent gain.	True	Conflicting
	"And while (Ted) Strickland proposed cuts for services for children, he wasted over $250,000 remodeling his bathrooms at the governor's mansion."	False	Conflicting
PASTA	She "led the fight to stop health insurance rate hikes and saved Rhode Island families over $150 million."	Conflicting	True
	406 constituencies voted to leave, 242 voted to remain in the EU referendum.	True	False
	Says U.S. Rep. Tammy Baldwin is backing President Barack Obama's plan to pass a $1.5 trillion tax increase.	False	True
Elastic-RoBERTa	Says Donald Trump's tax plan gives the wealthy and corporations "more than the Bush tax cuts by at least a factor of two."	True	Conflicting
	Said, "The Seven Years' Warled to near bankruptcy for many countries; Britain's need to raise taxes fueled the American desire for independence."	True	Conflicting
	Marylyn Addo habe bei einem Vortrag gesagt, dass nach der ersten Dosis des Impfstoffes der Firmen Biontech und Pfizer 20 Prozent der Studienteilnehmer Fieber bekamen. Nach der zweiten Dosis hätten 60 Prozent an Schüttelfrost gelitten.	Conflicting	False

Table 3: Randomly sampled incorrect claims for each general and numerically inclined model at each trickiness level.

Decomposition Method	Pre-trained				Finetuned			
	Rouge1	Rouge2	RougeL	RougeLSum	Rouge1	Rouge2	RougeL	RougeLSum
CLAIMDECOMP	33.92	17.07	28.58	31.05	40.70	20.99	35.80	37.12

Table 4: Evaluation of the generative Flan-T5-Base model finetuned on the CLAIMDECOMP dataset to decompose claims. Note that the Pseudo Program-FC method is not evaluated as we have no ground truth to compare our predictions to. Note that our custom decomposition method is based on the CLAIMDECOMP method, so there is no separate evaluation for this method in this step.

Figure 3: ROUGE curves during the training of the generative Flan-T5-Base model fine-tuned on the CLAIMDECOMP dataset.

Decomposition Method	Claim	Decomposed Questions
CLAIMDECOMP	"5.7 million – that's how many illegal immigrants might have voted" in 2008.	1. "Did 5.7 million illegal immigrants vote in 2008?" 2. "Are there any reasons why illegal immigrants didn't vote?" 3. "How many illegal immigrants may have voted in 2008?"
Pseudo Program-FC	"5.7 million – that's how many illegal immigrants might have voted" in 2008.	1. "Did 5.7 million illegal immigrants vote in the 2008 election?"
Custom (Statistical)	"5.7 million – that's how many illegal immigrants might have voted" in 2008.	1. - 3. [Same as 1. - 3. above] 4. What do the quantities 2008, 5.7 million mean in this claim? Are these quantities rigorously proven by the evidence?
Custom (Temporal)	Screenshots that surfaced online in early October 2021 authentically show photographs of, or comments written by, the Zodiac killer.	1. - (N-1). [From CLAIMDECOMP method] N. What do the dates early October 2021 mean in this claim? Are these dates rigorously proven by the evidence?
Custom (Interval)	Taco Bell is offering free gift cards valued at $60 and $75 on Facebook for the company's anniversary.	1. - (N-1). [From CLAIMDECOMP method] N. This claim contains a range or interval of numbers. What do the numbers 75, 60 mean in this claim? Are these numbers rigorously proven by the evidence?
Custom (Comparison)	The American Health Care Act will reduce premiums by 10 percent. We would actually bring costs down.	1. - (N-1). [From CLAIMDECOMP method] N. This claim contains a comparison of numbers. What do the numbers 10 percent mean in this claim? Are these numbers rigorously proven by the evidence?

Table 5: Sample decompositions of a randomly sampled QuanTemp claim for each decomposition method. For the custom decomposition method, the first N-1 sub-questions are generated by the CLAIMDECOMP method and are thus omitted from this table. The Nth sub-question is automatically added by the detected type of claim.

Model (Decomposition Method)	Statistical		Temporal		Interval		Comparison		Per-class F1			QuanTemp	
	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	M-F1	W-F1	T-F1	F-F1	C-F1	M-F1	W-F1
RoBERTa-Large-MNLI (Gold Standard)	54.08	60.46	63.39	79.40	55.44	67.73	53.57	53.69	47.11	82.17	42.34	57.21	65.96
RoBERTa-Large-MNLI (Original Claim)	**53.79**	**60.09**	**51.85**	**73.38**	**51.82**	**64.88**	**57.81**	**57.91**	38.46	**80.13**	**47.99**	**55.53**	**64.51**
RoBERTa-Large-MNLI (CLAIMDECOMP)	47.03	53.78	40.21	68.24	44.41	59.07	39.38	36.91	43.25	79.45	16.43	46.38	57.47
RoBERTa-Large-MNLI (Pseudo Program-FC)	41.19	47.93	45.42	69.14	40.36	54.69	33.1	30.23	**49.13**	76.32	1.32	42.25	53.17
RoBERTa-Large-MNLI (Custom Decomp)	51.38	57.67	37.25	66.6	38.39	57.17	54.22	55.08	32.8	78.71	40.55	50.69	60.84
BART-Large-MNLI (Gold Standard)	63.49	67.94	71.38	83.15	64.79	72.76	59.46	58.31	64.35	84.58	47.37	65.43	71.82
BART-Large-MNLI (Original Claim)	54.13	57.99	62.35	77.76	**54.20**	63.88	50.25	49.49	51.91	76.96	41.70	56.86	63.75
BART-Large-MNLI (CLAIMDECOMP)	**58.51**	**62.39**	**63.31**	**78.16**	50.29	61.39	53.41	52.92	**53.89**	78.45	45.03	**59.12**	**65.77**
BART-Large-MNLI (Pseudo Program-FC)	55.61	60.78	59.35	76.39	54.15	**65.74**	**55.59**	**55.95**	47.17	79.32	**47.09**	57.86	65.49
BART-Large-MNLI (Custom Decomp)	56.04	61.59	55.17	74.74	50.48	64.7	52.99	54.04	43.14	**80.81**	45.41	56.45	65.17
DeBERTa-v3-Base-Tasksource-NLI	63.83	68.15	67.17	80.99	63.28	73.51	60.97	60.16	62.91	84.81	47.33	65.02	71.67
DeBERTa-v3-Base-Tasksource-NLI (Original Claim)	54.59	60.06	57.28	75.51	46.62	61.25	50.8	49.98	48.78	**80.08**	35.91	54.92	63.55
DeBERTa-v3-Base-Tasksource-NLI (CLAIMDECOMP)	**57.03**	**61.38**	**57.42**	**75.71**	**52.01**	**64.32**	**58.25**	**58.08**	49.89	79.1	**46.62**	**58.54**	**65.76**
DeBERTa-v3-Base-Tasksource-NLI (Pseudo Program-FC)	55.15	60.15	55.47	74.45	44.61	59.67	50.02	48.58	**50.65**	79.49	33.62	54.59	63.02
DeBERTa-v3-Base-Tasksource-NLI (Custom Decomp)	52.19	57.61	55.7	75	44.6	59.97	52.48	51.69	49.5	79.25	31.39	53.38	62.13
Flan-T5-Base (Gold Standard)	58.88	63.13	68.84	81.12	59.19	67.39	57.41	56.70	58.56	80.69	45.57	61.61	68.07
Flan-T5-Base (Original Claim)	49.6	56.46	45.46	70.19	41.66	57.91	52.08	53	30.79	78.25	**43.28**	50.77	60.85
Flan-T5-Base (CLAIMDECOMP)	48.01	55.76	41.63	68.93	43.64	59.76	52.31	53.27	26.27	79.27	42.86	49.46	60.47
Flan-T5-Base (Pseudo Program-FC)	48.99	56.27	**46.11**	**70.79**	**47.61**	**62.36**	**54.98**	**55.59**	31.93	**79.55**	42.63	51.37	61.65
Flan-T5-Base (Custom Decomp)	**53.08**	**58.66**	46.07	70.59	44.72	59.68	51.52	50.28	**46.11**	79.19	33.71	**53.00**	**62.00**
Math-RoBERTa (Gold Standard)	59.08	64.28	66.31	79.91	58.18	68.18	60.76	60.30	56.13	82.58	46.15	61.62	62.82
Math-RoBERTa (Original Claim)	53.27	58.86	46.14	70.18	45.53	60.67	51.35	50.89	42.08	78.78	37.87	52.91	62
Math-RoBERTa (CLAIMDECOMP)	51.99	58.02	43.98	69.52	45.33	60.89	49.23	48.98	38.12	78.87	38.72	51.9	61.5
Math-RoBERTa (Pseudo Program-FC)	54.01	59.31	47.93	71.22	45.89	60.46	50.8	50.3	45.04	78.83	37.72	53.86	62.56
Math-RoBERTa (Custom Decomp)	**54.71**	**59.95**	**49.9**	**72.17**	**49.79**	**63.49**	**52.63**	**52.24**	**46.12**	**79.24**	**40.47**	**55.28**	**63.66**
Numeric-T5 (Gold Standard)	55.81	64.46	55.00	75.29	53.46	65.11	51.01	50.94	45.60	80.45	41.37	55.81	64.46
Numeric-T5 (Original Claim)	40.78	49.33	34.77	65.35	39.78	56.77	45.34	47.63	7.66	75.95	**44.64**	42.75	55.47
Numeric-T5 (CLAIMDECOMP)	**48.00**	**54.49**	**39.88**	**67.82**	44.74	59.69	51.37	**51.34**	**33.21**	**77.41**	37.72	**49.45**	**59.50**
Numeric-T5 (Pseudo Program-FC)	46.87	53.65	36.41	66.07	**46.18**	**61.13**	**51.96**	**51.34**	32.05	77.35	36.65	48.68	58.99
Numeric-T5 (Custom Decomp)	46.63	53.77	37.74	66.62	41.66	58	47.71	48.23	24.3	77.05	40.86	47.4	58.35
PASTA (Gold Standard)	51.16	57.24	42.39	69.50	46.07	61.13	52.16	50.76	48.20	80.81	23.68	50.90	60.92
PASTA (Original Claim)	51.80	57.58	**60.72**	**76.78**	49.96	63.11	56.06	56.11	**44.34**	**78.47**	43.87	55.56	63.70
PASTA (CLAIMDECOMP)	46.99	54.29	32.53	64.43	39.67	58.4	48.08	49.5	23.14	78.08	38.62	46.61	58.18
PASTA (Pseudo Program-FC)	**53.48**	**58.76**	59.32	76.26	**50.39**	**63.56**	**56.64**	**57.12**	42.7	78.45	**48.21**	**56.45**	**64.41**
PASTA (Custom Decomp)	47.46	54.64	33.19	64.72	41.44	59.38	48.41	49.79	24.28	78.18	39.28	47.24	58.61
Elastic-RoBERTa (Gold Standard)	63.65	68.04	72.16	83.00	64.94	72.19	60.05	60.55	58.72	83.52	60.87	65.78	72.00
Elastic-RoBERTa (Original Claim)	**55.27**	**59.42**	61.06	77.33	**58.83**	**66.74**	50.70	50.22	**52.40**	77.59	43.71	**57.90**	**64.68**
Elastic-RoBERTa (CLAIMDECOMP)	54.11	58.51	**62.62**	**77.92**	50.16	61.35	48.96	47.65	50.33	78.08	38.45	55.65	63.31
Elastic-RoBERTa (Pseudo Program-FC)	54.11	59.30	48.99	71.97	50.84	63.48	49.58	48.35	49.62	79.80	31.99	53.80	62.61
Elastic-RoBERTa (Custom Decomp)	52.44	58.81	54.5	74.36	52.45	65.43	**57.48**	**57.63**	41.33	**80.01**	**45.04**	55.46	64.28

Table 6: Results evaluating different claim decomposition methods on fine-tuned test models on the QuanTemp dataset.

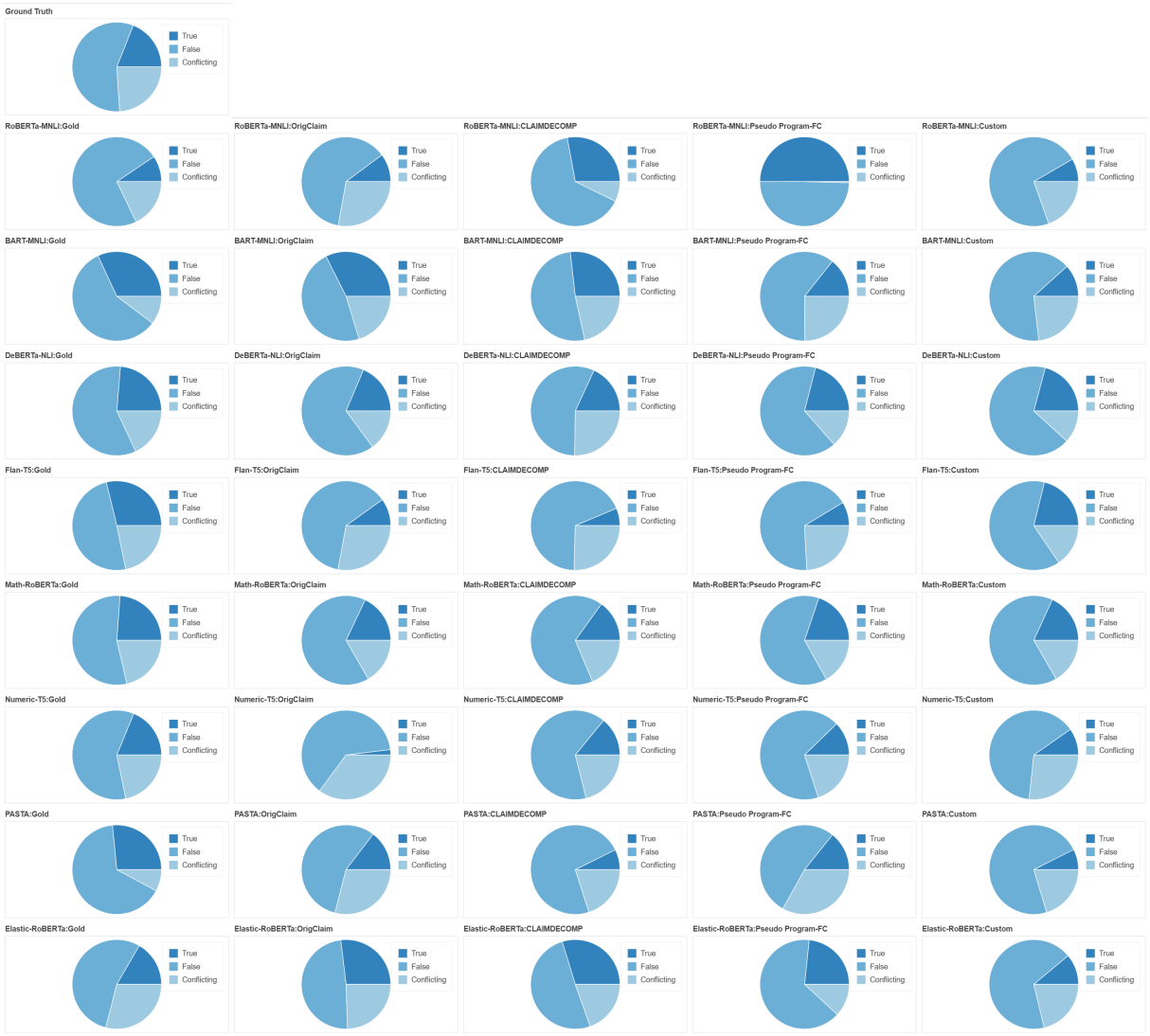

Figure 4: Pie charts of the distribution of predicted labels for each model fine-tuned on sub-questions and evidences generated by different claim decomposition methods on the QuanTemp test set. Ground truth pie chart also included for comparison.

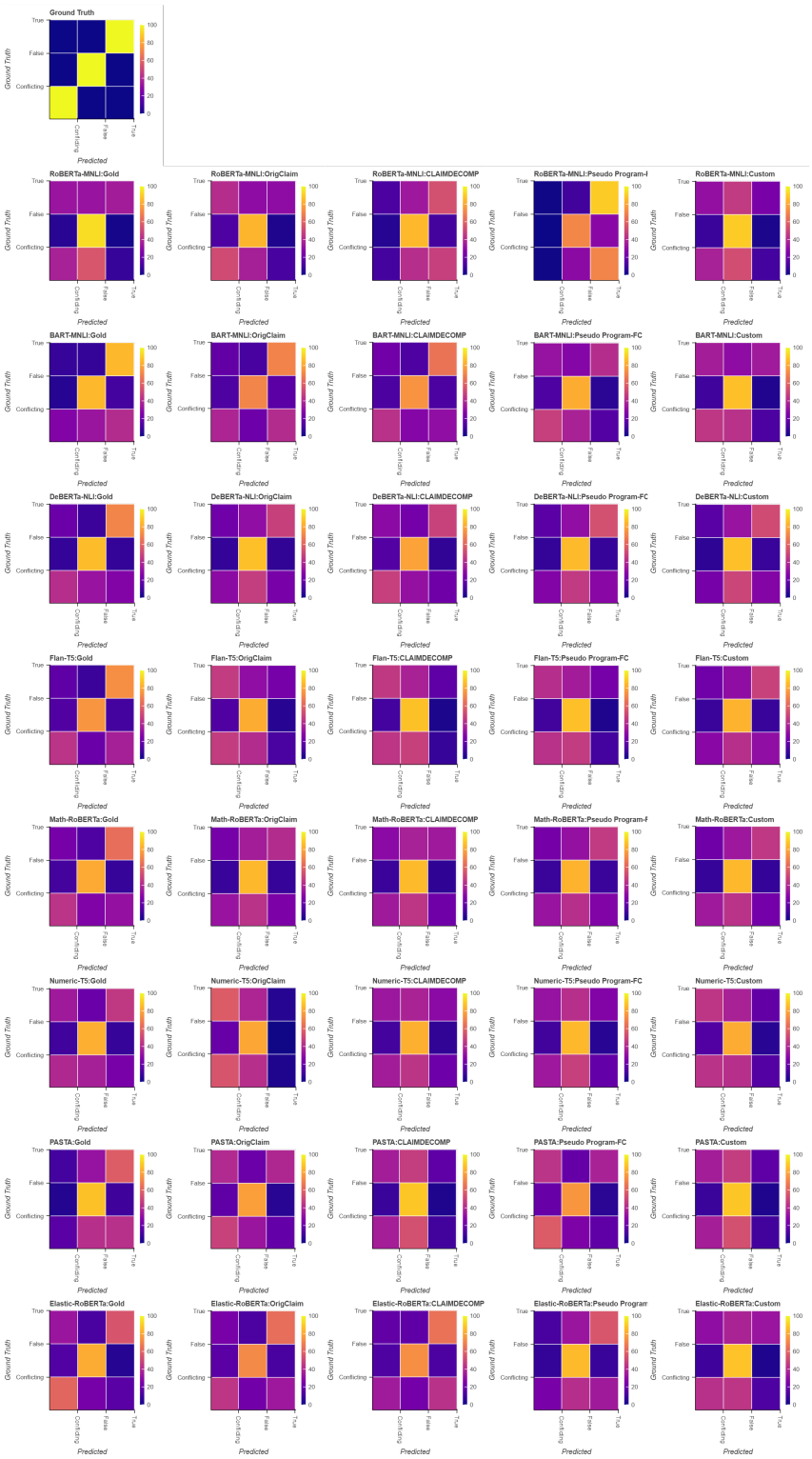

Figure 5: Confusion matrices for each model fine-tuned on sub-questions and evidences generated by different claim decomposition methods on the QuanTemp test set. Ground truth confusion matrix also included for comparison. Values in the confusion matrix go from 0 to 100%, signifying that 100% of the samples truly in that class were predicted as such.