Enhancing Mask Predictions for Text Anonymization

Derrick Chan-Sew
dchansew@berkeley.edu

Jason Dong
djliang@berkeley.edu

Abstract

As the demand for privacy-preserving technologies grow, effective identification of personal identifiable information (PII) has become a critical challenge across various domains and regulations. Traditional named entity recognition (NER) and existing datasets fall short in accurately identifying both direct and indirect personal identifiers required for text anonymization. In this work, we explore novel approaches to enhance token predictions by leveraging entity information for token classification. Our experiments demonstrate a consistent improvement of 4% and 6% on recall and F1-score against baselines, highlighting the effectiveness of attention mask manipulation using known entity information.

1 Introduction

Text anonymization of personal identifiable information (PII) is a crucial task spanning from healthcare to law as data is generated and released into the public sphere. With an overwhelming amount of data being scoured to train state-of-the-art language models, text anonymization is becoming increasingly important as individuals become more educated on privacy. The regulatory environment is also adapting to a data-rich landscape with an increased focus on consumer protections and privacy. Regulations such as the CCPA [6], HIPAA [1] and GDPR [9] are evolving with the rapid pace of technological change.

With the proliferation of large language models readily accessible to the public, preventing PII leakage from inference attacks is also becoming critical. C4, one of the largest language datasets used for foundation models, contains millions of non-anonymized personal information [16]. Other research has shown that various prompt injection attacks such as extraction, reconstruction, and inference can be highly successful in extracting masked personal information [7]. Techniques such as differential privacy [2] [4] and text anonymization provide safeguards but come with potential tradeoffs in model performance and utility [16].

Datasets such as MIMIC, CoNLL, and WikiPII have been used extensively for named entity recognition (NER) and by proxy PII detection. These datasets assume that all NER entities need to be masked and typically only identify 'Direct' identifiers which can directly disclose an individual's identity. Text anonymization, defined as the task of editing a document to prevent the disclosure of personal information, requires a broader approach. 'Quasi', aka indirect, identifiers including a person's appearance, profession, or religion can also lead to PII disclosure. For the purposes of text anonymization these direct and indirect identifiers must both be masked. The Text Anonymization Benchmark (TAB) is a novel dataset based on court cases from the European Court of Human Rights specifically designed for the broader anonymization task, addressing the narrow focus of de-identification by previous datasets and models.

In this work, we use TAB to predict the direct and indirect masking requirements of text spans for the broader goal of text anonymization. We start with a Longformer model capable of ingesting longer inputs and take inspiration from previous works adjusting positional [15] and word embeddings [17] [8] to pass entity information to enhance classification. We show that, using entity information, simple adjustments to the attention mask can nudge the model towards better class predictions.

2 Background

2.1 Task and Previous Work

The definition of personal information varies, often with vague and conflicting definitions. Recent sequence labeling techniques focus on de-identification and removing information that directly identifies a subject, but may miss more nuanced information such as physical appearance, profession or political opinions [14]. Previous evaluation metrics focused purely on these direct identifiers tend to overestimate the privacy protections of their models. As defined by GDPR, anonymization requires removing or masking any information that individually directly or in aggregate indirectly may re-identify the subject [10]. We build on this previous work by investigating and seeking to improve TAB's baseline implementation aimed at anonymization.

The various methods to anonymize documents can be distilled into two stages, identification and anonymization. During identification, text spans, co-reference information, and other attributes are produced for downstream task. Anonymization can occur in various methods with the most widely adopted techniques being Removal, Categorization, or Pseudonymization as summarized in Table 1. Presidio, a publicly available text anonymization module designed by Microsoft, allows individuals to adjust the model underlying the identification stage and subsequent anonymization methods [12].

Identification	**Anonymization**
Direct identifiers	Removal
Quasi-identifiers	Categorization
Masking	Pseudonymization
Co-reference	

Table 1: Identification vs. Anonymization

PII detection techniques have evolved over time following the progression of research starting from systematic expressions and dictionary based models [11] to CRFs, LSTMs, and transformer based models. While traditional models, including Presidio, show strong performance, custom transformer models still stand out on the text anonymization front [3] The BERT family of models output token-level embeddings making them particularly suitable for entity-level tasks. With 95% of our data exceeding the 512 token limit of BERT and RoBERTa, we proceeded with Longformer which builds on top of RoBERTa by adding a local attention window to enable processing longer inputs of text. [5].

3 Methods

3.1 Dataset

The TAB dataset comprises of 1,268 court cases developed in multiple stages starting with identifying PII text spans, determining masking requirements of Direct, Quasi, or No Mask for each span, and enriching annotations with additional information such as type of confidentiality and co-reference needed for the end-to-end anonymization [14]. TAB uses similar categories as traditional named entity recognition, but expands the spans identified in prior datasets. A separate CODE entity type is present, corresponding to court case numbers unique to the legal domain.

274 court cases contain unreconciled masking classification from multiple annotators. The TAB paper passes these documents multiple times during training and inference, once per annotator. We exclude these court cases from our experimentation to maintain ground truth. Text spans were split into word tokens using spaCy with the corresponding labels further processed to match IOB formatting. We then applied wordpiece tokenization and sequence padding to allow for ingestion into Longformer. The distribution of entities and masking decisions of our resulting dataset containing 994 court cases is shown in Table 2. Unless otherwise noted, our experiments were performed on a smaller subset of 400/50/50 court documents due to resource capacity.

3.2 Experiments

We use the HuggingFace transformers implementation of Longformer including a dropout and classification layer as our basis token classification. We evaluate this model against spaCy and Presidio with the assumption that all identified entities require masking. We convert our dataset into a binary MASK (Direct + Quasi) vs. NO MASK to follow these baseline assumptions.

We utilize the entity types and their corresponding positions to update the attention mask prior to ingestiong by Longformer. Despite documentation indicating that the attention mask inputs are restricted to $\{0, 1\}$, the attention layers in the HuggingFace implementation do not restrict cal-

Entity Type	Mentions	%	Direct	Direct %	Quasi	Quasi %	No Mask
DATETIME	29502	0.35	7	0.00	26005	0.88	3490
ORG	24048	0.28	10	0.00	8691	0.36	15347
PERSON	13145	0.16	2123	0.16	8323	0.63	2699
LOC	5251	0.06	1	0.00	3833	0.73	1417
DEM	4499	0.05	1	0.00	2128	0.47	2370
MISC	3612	0.04	22	0.01	2275	0.63	1315
CODE	2138	0.03	1186	0.55	758	0.35	194
QUANTITY	2218	0.03	0	0.00	1843	0.83	375
Total	84413	1.00	3350	0.72	53856	4.89	27207

Table 2: Distribution of entity types and masking decision after removing duplicate annotations

culations to those discrete values. We apply two methods for adjusting our attention masks, first by overweighting entities requiring masking decisions and underweighting where only O tokens while holding other attention weights constant.

For our concatenation model, we fine-tuned a separate Longformer model for traditional NER on entity types in TAB, took the resulting embeddings, and concatenated them with the embeddings from our best performing base model resulting in a shape of (4096, 1536) per document. Our neural network consisted of between 1 to 3 hidden and dropout layers with hidden dimensions of $\{512, 128\}$ and rate of 0.1 respectively before the final classification layer. Due to computation limitations, our discussion for these results are in Appendix D.

Unless specified, we used our mini subset of 400/50/50 to train Longformer. We explored various techniques to reduce the memory footprint of Longformer and our dataset with our limited computing resources. After various engineering efforts, we landed on a Pytorch implementation using 16 bit floating point precision and gradient checkpointing that could fit within a single NVIDIA T4 GPU. We fine-tuned Longformer with hyperparameters suggested by [5] and [13] as seen in Table 3 with training times ranging between 45 min - 2.5 hours.

3.3 Metrics

There is currently no agreed-upon system for evaluating text anonymization, thus we use SeqEval's metrics on text spans to evaluate our model [16]. We selected recall as our primary metric as false negatives would lead to PII disclosure, thus it would be better if our model was cautious in making its masking determination. We use PyPi's SeqEval implementation in default mode which excludes O tokens and allows leniency if the entity type is appropriately predicted but differs in B or I categorization.

We also use a benchmark provided by the TAB paper for direct comparisons against the original paper. These benchmarks utilize recall metrics specific for their dataset which can be seen in Appendix B.

These metrics reflect the degree of privacy protection. We consider that an entity is correctly masked if and only if the anonymization model manages to completely mask all of its mentions. If that condition is not met, the entity is counted as a false negative. [14]

4 Results and Discussion

Our best performing models for each experiment can be seen in Table 4. We are unable to produce the same baselines as TAB paper due to our different treatments of duplicate annotations. Our fine-tuned Longformer model performs better than pre-trained spaCy and Presidio models. The results do not come as a surprise due to differences in entity types and the assumption that all identified entities are a direct mask for base models. Changing our labels to binary prediction increased the SeqEval scores but the overall TAB benchmark remains the same.

4.1 SeqEval vs. TAB benchmark

We notice an interesting pattern where the SeqEval metric changes with our experimentation, while the TAB benchmark remains stagnant. Upon inspection, we find a couple of key discrepancies between the metric calculation. First, the TAB benchmark only checks if the model accurately predicts if a span should be masked. Only upon

Parameters	Values
Learning rate scheduler	Linear, Linear w/ warmup, Cosine w/ warmup
Learning rates	5e-4, 1e-4, 2.5e-5,
Warmup ratio	10%
Batch size	16, 8
Epochs	20
Early Stopping	3

Table 3: Hyper-parameters selected for fine-tuning our best performing models. We find that a learning rate = 2.5e-5 and scheduler = linear w/ warmup yielded the best results. A batch size = 8 was only used to reduce memory usage during training when needed.

Model	SeqEval				TAB	
	Train Recall	Test F1	Train Recall	Test F1	Train Recall	Test Recall
spaCy						0.88
Presidio						0.70
Longformer	0.79	0.75	0.72	0.71	0.96	0.94
Longformer Binary	0.77	0.75	0.74	0.74	0.96	0.94
Attention (overweight)	0.89	0.08	0.80	0.07	0.96	0.94
Attention (underweight)	0.83	0.81	0.83	0.81	0.96	0.94
Concatenated*	0.02	0.04	0.04	0.07	0.97	0.91

Table 4: Results from our training. The attention overweight model scales the attention mask by (1, 1.5) whereas the underweight model scales the attention mask by (0.75, 1). The concatenated model was only trained on 64 samples due to our limitation of computing resources.

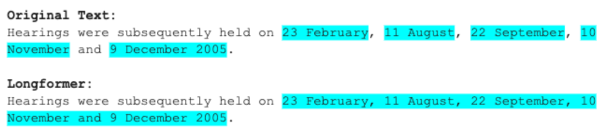

Figure 1: SeqEval considers this example as five text spans, leading to lower recall and precision.

ingestion into their gold standard does the TAB benchmark reclassify the spans as Direct vs. Quasi based on their source dataset. Second, the TAB benchmark is more lenient on span boundaries as we can see in Figure 2 where SeqEval punishes the prediction even though the text is anonymized appropriately.

4.2 Attention Mask Adjustments

Model performance for our attention masking experiments show an improvement compared to our baselines. Although overweighting the attention mask provides a higher recall score, we find that the improvement is negated by its performance on precision as seen by it's 0.08 F1-score. This result follows conventions to constrain values between (0, 1) to maintain numerical stability. We consistently observed higher recall F1-scores across all models where we underweighted the attention mask. Review of the SeqEval metrics shows differences in recall for the individual classes as seen in Table 5 of Appendix E. Our model with the best overall, Direct, and Quasi recall had attention weights of {0.75, 1} showing a improvement of 4% on recall and 6% on F1-scores against our baseline Longformer model. We see that the F1-scores are relatively consistent between the models thus further cross-validation is needed to verify the fluctuations for recall. We ran the test set with underweighted attention masks in our original base model and saw no differences in performance, validating the attention mask adjustments impact on model fine-tuning.

5 Conclusion

We present a novel approach to improving class prediction accuracy for text anonymization tasks by strategically underweighting the attention mask in transformer-based models. We demonstrate that this technique enhances the model's ability to differentiate between Direct and Quasi-identifiers, particularly in long-form legal texts where contextual understanding is paramount. Although the class distinction doesn't change the masking decision for text anonymization, we can see applications of this technique to any secondary objective where entity locations are known. This improvement highlights the potential of attention mask adjustments as a lightweight yet effective intervention for fine-tuning transformer models in complex entity recognition tasks. Moreover, we confirm that these adjustments influence the training phase without affecting the evaluation on unaltered test sets, validating their role in optimizing token-level predictions.

References

[1] Office for Civil Rights (OCR). *Summary of the HIPAA security rule*. Oct. 2022. URL: https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

[2] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS'16. ACM, Oct. 2016. DOI: 10.1145/2976749.2978318. URL: http://dx.doi.org/10.1145/2976749.2978318.

[3] Dimitris Asimopoulos et al. *Benchmarking Advanced Text Anonymisation Methods: A Comparative Study on Novel and Traditional Approaches*. 2024. arXiv: 2404.14465 [cs.CL]. URL: https://arxiv.org/abs/2404.14465.

[4] Rouzbeh Behnia et al. "EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy". In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, Nov. 2022, pp. 560–566. DOI: 10.1109/icdmw58026.2022.00078. URL: http://dx.doi.org/10.1109/ICDMW58026.2022.00078.

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.

[6] *California Consumer Privacy Act (CCPA) — oag.ca.gov*. https://oag.ca.gov/privacy/ccpa. [Accessed 09-12-2024].

[7] Nicholas Carlini et al. "Extracting Training Data from Large Language Models". In: *USENIX Security Symposium*. 2020. URL: https://api.semanticscholar.org/CorpusID:229156229.

[8] Yanru Dong et al. "A Fusion Model-Based Label Embedding and Self-Interaction Attention for Text Classification". In: *IEEE Access* 8 (2020), pp. 30548–30559. DOI: 10.1109/ACCESS.2019.2954985.

[9] Michelle Goddard. "The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact". In: *International Journal of Market Research* 59.6 (2017), pp. 703–705. DOI: 10.2501/IJMR-2017-050. eprint: https://doi.org/10.2501/IJMR-2017-050. URL: https://doi.org/10.2501/IJMR-2017-050.

[10] Mike Hintze. "Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency". In: *International Data Privacy Law* 8.1 (Dec. 2017), pp. 86–101. ISSN: 2044-3994. DOI: 10.1093/idpl/ipx020. eprint: https://academic.oup.com/idpl/article-pdf/8/1/86/24691426/ipx020.pdf. URL: https://doi.org/10.1093/idpl/ipx020.

[11] Yabing Liu et al. "Identifying Personal Information in Internet Traffic". In: *Proceedings of the 2015 ACM on Conference on Online Social Networks*. COSN '15. Palo Alto, California, USA: Association for Computing Machinery, 2015, pp. 59–70. ISBN: 9781450339513. DOI: 10.1145/2817946.2817947. URL: https://doi.org/10.1145/2817946.2817947.

[12] Microsoft. *Microsoft Presidio — microsoft.github.io*. https://microsoft.github.io/presidio/. [Accessed 09-12-2024].

[13] Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. *Efficient Classification of Long Documents Using Transformers*. 2022. arXiv: 2203.11258 [cs.CL].

URL: https://arxiv.org/abs/2203.11258.

[14] Ildikó Pilán et al. "The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization". In: *Computational Linguistics* 48.4 (Dec. 2022), pp. 1053–1101. ISSN: 0891-2017. DOI: 10.1162/coli_a_00458. eprint: https://direct.mit.edu/coli/article-pdf/48/4/1053/2062009/coli_a_00458.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00458.

[15] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. *Self-Attention with Relative Position Representations*. 2018. arXiv: 1803.02155 [cs.CL]. URL: https://arxiv.org/abs/1803.02155.

[16] Nishant Subramani et al. "Detecting Personal Information in Training Corpora: an Analysis". In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)* (2023). URL: https://api.semanticscholar.org/CorpusID:260063235.

[17] Xinyu Wang et al. *More Embeddings, Better Sequence Labelers?* 2021. arXiv: 2009.08330 [cs.CL]. URL: https://arxiv.org/abs/2009.08330.

6 Appendices

6.1 Appendix A - Supplementary Materials

The appendices are organized as follows. Section B details the TAB benchmark from [14]. In Section C, we provide textual examples between the source text and predictions impacting and discuss its impact on SeqEval. In Section D, we provide more details on the attention mask experiments. In Section F, we discuss our initial results of our concatenation experiments.

6.2 Appendix B - TAB Benchmark

Let D denote a set of documents, where each document $d \in D$ is represented as a sequence of tokens. Let A be a set of expert annotators, and $E_a(d)$ be the set of entities that were masked by annotator a in the document d. Each entity $e \in E_a(d)$ is itself defined as a list of token indices T_e where that entity e is mentioned in the document d (there might be several mentions of a given entity in a document). Then, assuming that an anonymization model outputs a set of word indices $M(d)$ to mask in the document d, we count each entity e as a true positive if $T_e \subset M(d)$, and a false negative otherwise. In other words, we consider that an entity is correctly masked if and only if the anonymization model manages to completely mask all of its mentions. If that condition is not met, the entity is counted as a false negative [14].

We use separate recall measures for the direct identifiers (such as full person names, case numbers, etc.) and the quasi-identifiers (dates, locations, etc.). This distinction gives us a more fined-grained measure of the anonymization quality, since a low recall on the direct identifiers corresponds to a failure of the anonymization process (as it implies that the person identity is disclosed), independently of the coverage of other types of identifiers. The set of identifiers $E_a(d)$ marked by annotator a in the document d is thus split into two disjoint sets: a set $E_a^{di}(d)$ for the direct identifiers and a set $E_a^{qi}(d)$ for the quasi-identifiers [14].

As noted above, a document may admit more than one anonymization solution. To account for this multiplicity, we compute the recall and precision as micro-averages over all annotators.

The entity-level recall on direct identifiers ER_{di} is defined as the micro-averaged recall over the entities defined as direct identifiers:

$$ER_{di} = \frac{\sum_{d \in D} \sum_{a \in A} \sum_{e \in E_a^{di}(d)} \mathbf{1}(T_e \subset M(d))}{\sum_{d \in D} \sum_{a \in A} |E_a^{di}(d)|} \quad (1)$$

The entity-level recall on quasi-identifiers ER_{qi} is defined similarly:

$$ER_{qi} = \frac{\sum_{d \in D} \sum_{a \in A} \sum_{e \in E_a^{qi}(d)} \mathbf{1}(T_e \subset M(d))}{\sum_{d \in D} \sum_{a \in A} \left|E_a^{qi}(d)\right|} \quad (2)$$

6.3 Appendix C - Error Analysis

We provide examples of general themes from our error analysis which impacts the resulting metrics for our baseline model. In general, due to our focus on recall, our model tends to be more conservative than original source annotations. Additionally, the difference in span boundaries predicted by our model compared to the source text negatively penalizes the SeqEval metric which requires an exact match. The color coding for our text annotation is detailed below in Figure 2.

Direct = Yellow
Quasi = Blue
No Mask = Green

Figure 2: Direct, Quasi and No Mask Legend

6.3.1 Example: Conservative Predictions

doc_id = 001-66929

Original Text:
During the hearings before the Izmir State Security Court, the applicants denied the statements they had made to the police and the investigating judge . On 25 March 1997 the Izmir State Security Court, composed of two civilian judges and a military judge, convicted the applicants as charged and sentenced the first applicant to three years and nine months' imprisonment and the second applicant to twelve years and six months' imprisonment . Neither the applicants nor their lawyers were present in this last hearing . The applicants appealed . On 12 November 1997 the Court of Cassation dismissed the applicants' appeal , upholding the Izmir State Security Court's reasoning and assessment of evidence . The decision , which was pronounced in the absence of the applicants and their lawyers, was deposited with the Registry of the Izmir State Security Court on 21 November 1997.

Longformer:
During the hearings before the Izmir State Security Court, the applicants denied the statements they had made to the police and the investigating judge . On 25 March 1997 the Izmir State Security Court, composed of two civilian judges and a military judge, convicted the applicants as charged and sentenced the first applicant to three years and nine months' imprisonment and the second applicant to twelve years and six months' imprisonment . Neither the applicants nor their lawyers were present in this last hearing . The applicants appealed . On 12 November 1997 the Court of Cassation dismissed the applicants' appeal , upholding the Izmir State Security Court's reasoning and assessment of evidence . The decision , which was pronounced in the absence of the applicants and their lawyers, was deposited with the Registry of the Izmir State Security Court on 21 November 1997.

Figure 3: Our model tends to be more cautious than the original annotations. In the example below, courts were identified as Quasi text requiring masking indicating the ambiguity in determining when locations are permissible.

6.3.2 Example: Quasi vs. Direct Misattribution

doc_id = 001-95382

Original Text:
The case originated in an application (no . 21377/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (" the Convention ") by a Turkish national , Mr Ahmet Kenan Er (" the applicant ") , on 30 April 2004 . The applicant was represented by Mr H.İ. Er , a lawyer practising in Istanbul.

Longformer:
The case originated in an application (no . 21377/04) against the Republic of Turkey lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms (" the Convention ") by a Turkish national , Mr Ahmet Kenan Er (" the applicant ") , on 30 April 2004 . The applicant was represented by Mr H.İ. Er , a lawyer practising in Istanbul.

Figure 4: We show misclassified Quasi and Direct identifiers for the applicant and lawyer negatively impacting the SeqEval metric. Although misclassified, both of the labels lead to masking and are not punished by the TAB benchmark.

6.3.3 Example: Annotation Quality

doc_id = 001-61177

Original Text:
of the Ministry of Foreign Affairs . \n\n The application was transmitted to the Court on 1 November 1998 , when Protocol No . 11 to the Convention came into force (Article 5 § 2 of Protocol No . 11)

Longformer:
of the Ministry of Foreign Affairs . \n\n The application was transmitted to the Court on 1 November 1998 , when Protocol No . 11 to the Convention came into force (Article 5 § 2 of Protocol No . 11)

Figure 5: Our model identifies spans not originally annotated from source text which may be beneficial for text anonymization.

6.4 Appendix D - Concatenation Discussion

Concatenating the embeddings for our models produced 9.8 GB of data. We were unable to successfully stream the data across the distributed files and utilized a smaller subset of 64 examples to stream from a single file. Despite the poor model performance on the SeqEval metrics, the model still had a .961 score from the TAB benchmark. Although the distinction between Direct vs. Quasi failed, the model still identifies those tokens and their need for masking. With embedding sizes increasing exponentially with input size, we do not believe our design is tenable without further dimension reduction. For future experimentation, we would add additional dense layers to reduce the dimensionality of the embeddings prior to concatenation to reduce the footprint of the dataset. Another approach could be averaging the embeddings along the token axis; however, we did not find existing literature exploring these techniques for token classification.

Experiment	Mask Adjustments {O, DQN}	Direct			Quasi			Overall		
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Longformer Base	{1, 1}	**0.70**	0.89	0.79	**0.82**	0.75	0.78	**0.72**	0.71	0.71
Attention (over)	{1, 1.5}	**0.73**	0.80	0.76	**0.57**	0.01	0.01	**0.80**	0.03	0.07
Attention (under)	{0.75, 1}	**0.70**	0.69	0.69	**0.91**	0.78	0.84	**0.79**	0.77	0.76
Attention (under)	{0.5, 1}	**0.68**	0.83	0.75	**0.50**	0.76	0.61	**0.80**	0.77	0.79
Attention (under)	{0.25, 1}	**0.59**	0.87	0.70	**0.92**	0.76	0.83	**0.78**	0.77	0.76

Table 5: Direct and Quasi SeqEval results from our training

6.5 Appendix E - Detailed SeqEval Metrics

We provide SeqEval metrics for our Direct and Quasi classes in Table 5. We selected attention mask weights $\{0.75, 0\}$ as our best performing model due to best recall for Direct and Quasi labels in addition to the high overall score. Further confirmation via cross-validation can provide insights behind the fluctuation.