An Exploration of Encoder-Decoder Approaches to Multi-Label Classification for Legal and Biomedical Text

Yova Kementchedjhieva* **Ilias Chalkidis***
Department of Computer Science, University of Copenhagen, Denmark
{yova,ilias.chalkidis}[at]di.ku.dk

Abstract

Standard methods for multi-label text classification largely rely on encoder-only pre-trained language models, whereas encoder-decoder models have proven more effective in other classification tasks. In this study, we compare four methods for multi-label classification, two based on an encoder only, and two based on an encoder-decoder. We carry out experiments on four datasets—two in the legal domain and two in the biomedical domain, each with two levels of label granularity— and always depart from the same pre-trained model, T5. Our results show that encoder-decoder methods outperform encoder-only methods, with a growing advantage on more complex datasets and labeling schemes of finer granularity. Using encoder-decoder models in a non-autoregressive fashion, in particular, yields the best performance overall, so we further study this approach through ablations to better understand its strengths.

1 Introduction

Multi-label classification constitutes the task of predicting multiple labels for an input as opposed to a single (possibly binary) one. The labels are drawn from a set of up to several hundred classes, often with the added challenge of class imbalance. While the order in which labels are predicted is irrelevant, there can be interdependence between subsets of labels. The task is commonly approached with a classification model based on a pre-trained encoder followed by a multi-output classification head.

Encoder-decoder models, like T5 (Raffel et al., 2020), have taken over recent NLP literature with state-of-the-art results on various tasks, such as question-answering (QA), summarization, single-label classification, etc. Raffel et al. (2020) showed that any given NLP task could be reformulated as a *text-to-text* task and solved with conditional generation, i.e., generating a text sequence that represents the desired output, be that a span of text in QA, a text summary, a label descriptor, etc. Liu et al. (2021) presented an alternative use of encoder-decoder models for classification tasks in particular, wherein T5's decoder is used in a non-autoregressive fashion to obtain output representations, which are then fed to a classification head.

The application of encoder-decoder methods to multi-label classification is currently limited to one experiment in the work of Liu et al. (2021), who compare a text-to-text approach and their non-autoregressive approach on a single dataset, including an encoder-only baseline built off of a different pre-trained model, BERT (Devlin et al., 2019). They obtain results favorable to the two encoder-decoder methods, but since the focus of their work is not multi-label classification in particular, their evaluation is insufficient to draw hard conclusions about this task, and analysis on the contribution of different model components to performance on the task is missing altogether.

In this work, we carry out an extensive study of encoder-decoder approaches to multi-label classification. To ensure the thorough and fair evaluation of all methods:

(a) We experiment on four datasets from two different domains (legal and biomedical), each with two levels of label granularity.

(b) We include four methods for multi-label classification, two encoder-only methods and two encoder-decoder methods.

(c) We conduct preliminary development to determine the best configuration for the application of each method, e.g. choice of label descriptors for the text-to-text approach.

(d) We explore how model size affects performance, by fine-tuning small, base, and large T5 models.

* Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 5828–5843
July 9-14, 2023

(e) We ablate components of the best performing approach, the non-autoregressive encoder-decoder method of Liu et al. (2021), to better understand its strengths.

We release our code base to assure reproducibility and let others extend our study by experimenting with new methods and more datasets.[1]

2 Related Work

Class imbalance is a critical issue in multi-label classification, with researchers searching for the best method to handle rare (less represented) labels.

Encoder-only Approaches Snell et al. (2017) introduced the idea of a *prototype* label vector, obtained by averaging over all instances of a given class and used to add inductive bias to their Prototypical Network for multi-label classification. In a similar vein, Mullenbach et al. (2018) developed the Label-Wise Attention Network (LWAN) architecture, in which label-wise document representations are obtained by learning to attend to the most informative input words for each label, using trainable label vectors as keys.

Chalkidis et al. (2020) systematically studied the effects of different language encoders (CNNs, BIGRUs, BERT) and several variants of LWAN with regards to the representation of prototype labels. Experimenting with three datasets (EURLEX, MIMIC-III, and AMAZON), they showed that better language encoders counter-play the positive effect of the LWAN module, i.e., a standard BIGRU classifier outperforms CNN-based LWANs (Mullenbach et al., 2018), and a standard BERT outperforms BIGRU-LWAN, respectively. Moreover, BERT-based LWANs offer minor overall improvements compared to a vanilla BERT classifier, wherein BERT's *CLS* token representation is passed to a classification head (Devlin et al., 2019).

Chalkidis et al. (2021) were the first to explore the use of a T5 model for multi-label classification, although they only considered an encoder-only classifier, disregarding the model's decoder. They followed the now standard approach of a classification head on top of the </s> token representation. In experiments with mT5 (Xue et al., 2021), they showcased improved results compared to XLM-R (Conneau et al., 2020) on a newly introduced multilingual dataset, MultiEURLEX.

[1] https://github.com/coastalcph/Multi-Label-Classification-T5

Encoder-Decoder Approaches Text-to-text approaches, which utilize the full encoder-decoder model, have proven effective for binary and single-label classification tasks (Raffel et al., 2020; Chung et al., 2022). The key to such approaches are label verbalizers, words in natural language which verbalize the underlying semantics of a given class. Label verbalizers are represented in the embedding space of pre-trained models and in this way benefit from the model pre-training. This can be more optimal especially for few- and zero-shot labels, in comparison to head-based classification methods where randomly initialized parameters have to be learned from scratch.

Liu et al. (2021) presented an alternative use of the full T5 model for non-autoregressive tasks, e.g. single-label and multi-label classification, wherein the decoder is used to obtain label-wise representations informed by the input document, which in turn are fed to label-specific binary classification heads. Liu et al. (2021) performed one set of experiments on the EURLEX-57K dataset (Chalkidis et al., 2019), in which they compared their non-autoregressive approach to a T5-based text-to-text approach and a standard BERT-based classifier. They found that both T5-based approaches outperformed the encoder-only classifier, the non-autoregressive method performing best. Nonetheless, the encoder-only classifier had less than half the parameters of the T5 model (110M vs 222M). Encoder-decoder approaches thus seem to carry potential for multi-label classification, still with insufficient empirical evidence, however.

3 Methods

We experiment with four methods for multi-label classification, *Encoder+Head*, *LWAN*, *Seq2Seq*, and *T5Enc*, basing their implementation on the T5 model (Raffel et al., 2020). T5 is a transformer-based encoder-decoder model (Vaswani et al., 2017), which encodes a string of input tokens and generates a string of output tokens.

All methods discussed below use T5's encoder to represent input documents, a document being denoted as $[x_1, x_2, \ldots, x_N]$, where N is the document length in terms of T5 subword tokens. Some methods further use the model's decoder—we introduce decoder notation where needed.

Encoder+Head In this case, we use only the encoder of T5 in the standard classification setting, as introduced by Devlin et al. (2019). We feed the

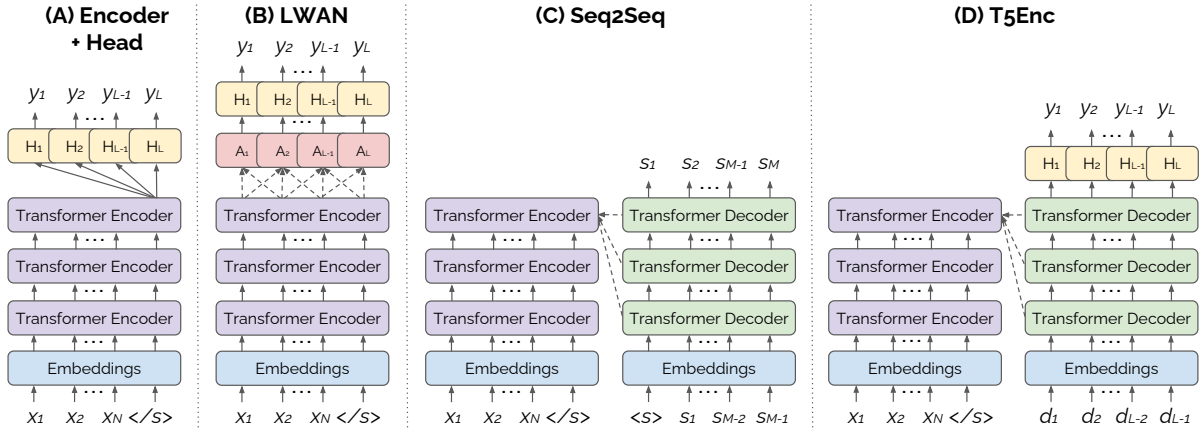


Figure 1: Depiction of the four task-specific methods for multi-label classification: encoder-only (*Encoder+Head*, *LWAN*), and encoder-decoder (*Seq2seq*, *T5Enc*). x: input tokens, y: label predictions, s: label descriptor tokens, d: label descriptors, N: input sequence length, L: label set size, M: length of tokenized label sequence.

document to the encoder, and use the representation of the special </s> token as document representation ($d \in \mathbb{R}^{dim}$). This representation is passed to L standard classification heads, one per label.

LWAN In this case, we use a Label-Wise Attention Network (LWAN) (Mullenbach et al., 2018) on top of the T5 encoder, as done in Chalkidis et al. (2020). We feed the document to the encoder, and use one attention head per label to generate L label-wise document representations $d_l \in \mathbb{R}^{dim}$, i.e., L weighted averages of the contextualized token representations. Intuitively, each head focuses on possibly different tokens of the document relevant to the corresponding label. LWAN employs L linear layers ($o_l \in \mathbb{R}^{dim \times 1}$) each operating on a different label-wise document representation d_l, to produce L scores (logits), one per label.

Seq2Seq In this case, we use T5 for conditional generation, which is the standard form of use, since T5 was trained in an autoregressive fashion. The target labels are formatted as a sequence of label descriptors, separated by a comma and a space, and ordered alphabetically, e.g., 'EU, finance'. We feed the document to the encoder and use the decoder to generate the tokenized output sequence, $[s_1, s_2, \ldots, s_M]$. When we evaluate the trained model's performance in inference time, we split the generated sequences using comma as a delimiter, keeping only valid label descriptors, and treat them as a set (since their order does mot matter for the task). We consider different options for the label descriptors, discussed in Section 5.2.

T5Enc In this case, we follow the work of Liu et al. (2021), where they use T5 in a non-autoregressive fashion.[2] We feed the document to the encoder, and use the decoder in non-autoregressive fashion, where its inputs are fixed (pre-populated), i.e., we feed the decoder with single-token label descriptors, $[d_1, d_2, ..., d_L]$, where L is the size of the full label set. We then use a binary classification head ($o_l \in \mathbb{R}^{dim \times 1}$) per decoder output representation to produce L scores, one per label. This method can be seen as an advanced version of the LWAN method which builds label-wise representations (d_l) via attention. In this case, however, these representations are further co-attended (conditioned) via the standard decoder self-attention across many decoder layers.

4 Datasets

We experiment with four datasets from the legal and biomedical domains, each with two different label granularities, i.e., label sets including more abstract or more specialized concepts.

UKLEX United Kingdom (UK) legislation is publicly available as part of the United Kingdom's National Archives.[3] Most of the laws have been categorized in thematic categories (e.g., healthcare, finance, education, transportation, planing), which are stated in the document preamble and are used for archival indexing purposes. The UKLEX dataset (Chalkidis and Søgaard, 2022) comprises

[2] We keep the name T5Enc, as coined by the authors, for consistency, although the model actually uses both the encoder and the decoder of T5.

[3] https://www.legislation.gov.uk/

36.5k UK laws. The dataset is chronologically split in training (20k, 1975–2002), development (8k, 2002–2008), and test (8.5k, 2008–2018) sets.

EURLEX European Union (EU) legislation is published on the EUR-Lex website. All EU laws are annotated by EU's Publications Office with multiple concepts from EuroVoc, a thesaurus maintained by the Publications Office.[4] EuroVoc has been used to index documents in systems of EU institutions. We use the English part of the dataset of Chalkidis et al. (2021), which comprises 65k EU laws (documents). The dataset is chronologically split in training (55k, 1958–2010), development (5k, 2010–2012), and test (5k, 2012–2016) sets. It supports four different label granularities. We use the 1st and 2nd level of the EuroVoc taxonomy.

BIOASQ The BIOASQ (Task A) dataset consist of biomedical articles from PubMed,[5] annotated with concepts from the Medical Subject Headings (MeSH) taxonomy (Tsatsaronis et al., 2015; Nentidis et al., 2021).[6] MeSH is a hierarchically-organized vocabulary produced by the National Library of Medicine. The current version of MeSH contains more than 29k concepts referring to various aspects of the biomedical research (e.g., diseases, chemicals and drugs). It is primarily used for indexing, cataloging, and searching of biomedical and health-related information. We subsample 100k documents from the period 2000-2021 in the latest version (v.2022) of the dataset, and split those chronologically for training (80k, 1964–2015), development (10k, 2015–2018), and testing (10k, 2018–2020). We use the 1st and 2nd levels of the MeSH taxonomy.

MIMIC-III The MIMIC-III dataset (Johnson et al., 2017) contains approximately 50k discharge summaries from US hospitals. Each summary is annotated with one or more codes (labels) from the ICD-9 hierarchy, which has eight levels in total.[7]. The International Classification of Diseases, Ninth Revision (ICD-9) is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. Documents in MIMIC-III have been anonymized to protect patient privacy, including chronological information (e.g., entry/discharge dates). Hence, it is not possible to split the data chronologically, so we split it randomly in train (30k), development (10k), and test (10k) sets. We use the 1st and 2nd level of the ICD-9 hierarchy.

Dataset	Size	|L1|	L/D	T/L	|L2|	L/D	T/L
UKLEX	36.5k	18	1.2	2.1	69	1.5	1.7
EURLEX	65k	21	3.2	2.4	127	4.5	2.9
BIOASQ	100k	16	5.6	3.4	116	8.9	4.0
MIMIC-III	50k	19	6.0	7.8	184	10.1	8.4

Table 1: Summary of datasets in terms of size, number of labels on Level 1 (|L1|) and 2 (|L2|), average number of gold labels per document (L/D), and average number of tokens per label (T/L) in the T5 vocabulary.

All four datasets come with label descriptors, e.g. 'Agriculture & Food', 'Immigration & Citizenship' (UKLEX), and 'Chemicals and Drugs', 'Skin and Connective Tissue Diseases' (BIOASQ).[8] More details about the datasets are provided in Table 1. Notice that Level 2 label sets are considerably larger than Level 1 label sets, and that the number of label assignments per document do not grow proportionately from Level 1 to Level 2, which means Level 2 labels have less representation on average.

5 Experiments

5.1 Experimental Setup

We use the original checkpoints of T5 released by Raffel et al. (2020) from the Hugging Face Hub.[9] Following Raffel et al., for all four methods we use the Adafactor optimizer (Shazeer and Stern, 2018) with a fixed learning rate of 1e-4 after warm-up for one epoch.[10] Seq2Seq models are trained with teacher forcing. We report results in terms of micro-F1 (μ-F_1), and macro-F1 (m-F_1) scores, the former more indicative of performance on well-represented labels, the latter, of performance on rare labels. When fine-tuning models, we use early stopping based on validation micro-F1 scores. We run each experiment with 4 seeds, and report the mean and standard deviations across runs.

5.2 Preliminary Experiments

We conduct a series of preliminary experiments to identify the most promising setting for the examined methods. All results reported here are on the development split of respective datasets.

[4] http://eurovoc.europa.eu/
[5] https://pubmed.ncbi.nlm.nih.gov
[6] https://www.nlm.nih.gov/mesh/
[7] www.who.int/classifications/icd/en/

[8] See Appendix B for label descriptors across all datasets.
[9] https://huggingface.co/t5-base
[10] In preliminary experiments, we also considered the widely used AdamW optimizer (Loshchilov and Hutter, 2017), which led to lower performance in most cases.

No. Heads	UKLEX (L1)		EURLEX (L2)	
	μ-F_1	m-F_1	μ-F_1	m-F_1
N=1	**83.3** ± **0.2**	**79.3** ± **0.7**	**76.3** ± **0.3**	**55.5** ± **0.8**
N=4	82.8 ± 0.3	78.1 ± 0.7	75.1 ± 0.1	51.7 ± 2.1
N=6	83.2 ± 0.3	**79.3** ± **0.5**	75.1 ± 0.3	54.1 ± 0.6
N=12	83.0 ± 0.4	78.8 ± 1.4	75.2 ± 0.3	53.0 ± 1.2

Table 2: Number of attentions heads for LWAN.

LWAN – Number of attention heads Previous work which employed the LWAN approach always used a single attention head in the label-wise attention mechanism. Here, we experiments with $N \in [1, 4, 6, 12]$. In Table 2, we reports results on two datasets, UKLEX (L1) with 18 labels, and EURLEX (L2) with 127 labels. We observe that in the case of UKLEX (L1) increasing the number of attention heads does not improve results, while in the case of EURLEX (L2) it harms performance. It appears that the added expressivity from multi-head attention is either not needed, or it is not easily utilized, since it adds more randomly initialized parameters which have to be learned from scratch. In subsequent experiments, we thus use the standard single-head attention mechanism.

Label	UKLEX (L1)		MIMIC (L1)	
	μ-F_1	m-F_1	μ-F_1	m-F_1
Original	84.2 ± 0.0	**81.6** ± **0.2**	73.2 ± 0.0	**70.2** ± **0.2**
Simplified	**84.8** ± **0.2**	78.7 ± 0.3	73.1 ± 0.1	70.1 ± 0.1
Numbers	83.8 ± 0.2	80.2 ± 0.7	**73.3** ± **0.1**	69.7 ± 0.2

Table 3: Form of label descriptors for Seq2Seq.

Seq2Seq – Form of Label Descriptors We consider three alternative forms of label descriptors:

(a) the *original* label descriptors, which may include complex multi-word expressions, e.g., 'Anthropology, Education, Sociology, and Social Phenomena'

(b) *simplified* versions of the original label descriptors, manually curated to consist of single-token expressions (as per the T5 vocabulary), e.g., 'Anthropology' for the example above

(c) *numbers* arbitrarily assigned to labels, e.g. '1'.

In Table 3, we present results on two datasets, UKLEX (L1), where the original label descriptors are mostly single-word expressions that map onto T5 sub-word tokens, and MIMIC (L1), where the original label descriptors are multi-word expressions which are further tokenized into subwords

We observe mixed rankings between the three forms of label descriptors across different metrics and datasets, with slight advantage for a lexical form over the arbitrary numerical one. This is in line with the intuition that the semantics of the label descriptors contribute to the learning of the task. In subsequent experiments, we use the original label descriptors across all datasets.

Decoding	UKLEX (L1)		MIMIC (L1)	
	μ-F_1	m-F_1	μ-F_1	m-F_1
Greedy	**84.3** ± **0.0**	**81.6** ± **0.2**	72.9 ± 0.2	69.4 ± 0.4
Beam	84.2 ± 0.0	**81.6** ± **0.2**	**73.2** ± **0.1**	**70.3** ± **0.2**

Table 4: Greedy decoding vs. beam search for Seq2Seq.

Seq2Seq – Greedy Decoding vs. Beam Search Raffel et al. (2020) suggested using greedy decoding for single-label classification tasks but also found beam search decoding (N=4) to work better for tasks with long output sequences, as is the case in multi-label classification. In Table 4, we compare the two decoding strategies on UKLEX (L1) and MIMIC (L1). We find that the choice of decoding strategy has little effect on performance, likely because the output space in these tasks is constrained to a fixed set of valid labels, in a single permissible (alphabetical) order. In subsequent experiments, we use beam search (N=4), as it performs slightly better on average.

Label	UKLEX (L1)		MIMIC (L1)	
	μ-F_1	m-F_1	μ-F_1	m-F_1
Simplified	**84.8** ± **0.2**	81.9 ± 0.5	**73.6** ± **0.2**	**69.2** ± **1.5**
Pseudo	**84.8** ± **0.1**	**82.3** ± **0.2**	73.2 ± 0.1	67.7 ± 1.9

Table 5: Form of label descriptors for T5Enc.

T5Enc – Form of Label Descriptors We compare two forms of label tokens, lexical (using simplified descriptors, as they have to be single tokens), and pseudo descriptors, where we introduce special tokens to the vocabulary of T5 (e.g., <label_1>). Results on UKLEX (L1) and MIMIC (L1) are presented in Table 5. We observe that results are comparable for UKLEX, while simplified label descriptors perform slightly better for MIMIC. In subsequent experiments, we thus use simplified label descriptors for Level 1 datasets. For Level 2 datasets,

Method	UKLEX (L1)		EURLEX (L1)		BIOASQ (L1)		MIMIC (L1)		Average	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
Enc+Head	**80.8** ± **0.5**	**77.2** ± **0.4**	78.9 ± 0.4	67.9 ± 1.1	86.4 ± 0.0	76.8 ± 0.1	72.2 ± 0.2	66.3 ± 0.7	79.6	72.1
LWAN	80.4 ± 0.3	76.6 ± 0.5	79.6 ± 0.4	68.4 ± 0.7	86.3 ± 0.1	77.2 ± 0.2	72.3 ± 0.3	66.8 ± 0.8	79.7	72.3
Seq2Seq	79.6 ± 0.6	76.4 ± 0.6	78.8 ± 0.2	69.1 ± 0.3	86.0 ± 0.1	77.8 ± 0.2	72.9 ± 0.1	**69.7** ± **0.2**	79.3	73.3
T5Enc	**80.8** ± **0.4**	77.1 ± 0.5	**80.0** ± **0.3**	**70.5** ± **0.4**	**86.6** ± **0.0**	**77.9** ± **0.4**	**73.4** ± **0.3**	68.8 ± 1.4	**80.2**	**73.6**
Method	**UKLEX (L2)**		**EURLEX (L2)**		**BIOASQ (L2)**		**MIMIC (L2)**		**Average**	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
Enc+Head	75.9 ± 0.5	64.9 ± 0.5	70.3 ± 0.2	48.2 ± 1.2	73.1 ± 0.0	60.1 ± 0.8	56.7 ± 0.6	22.3 ± 1.2	69.0	48.9
LWAN	**76.6** ± **0.2**	65.0 ± 0.8	70.3 ± 0.3	49.0 ± 0.7	73.0 ± 0.1	59.7 ± 0.9	57.2 ± 0.4	24.2 ± 0.3	69.3	49.5
Seq2Seq	75.3 ± 0.2	65.8 ± 0.4	70.6 ± 0.3	51.8 ± 1.0	73.8 ± 0.1	63.8 ± 0.1	57.4 ± 0.2	**31.2** ± **1.7**	69.3	53.2
T5Enc	76.5 ± 0.3	**66.8** ± **0.9**	**72.0** ± **0.2**	**53.2** ± **1.4**	**75.1** ± **0.1**	**66.0** ± **0.1**	**60.5** ± **0.1**	31.1 ± 0.9	**71.0**	**54.3**

Table 6: Test results for encoder-only methods (Encoder+Head and LWAN) and encoder-decoder methods (Seq2Seq and T5Enc) trained from T5-Base.

we use pseudo labels, since we cannot manually curate simplified descriptors for hundreds of labels.

Encoder	UKLEX (L1)		BIOASQ (L2)	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
BERT	**84.4** ± 0.3	**81.3** ± 0.9	71.7 ± 0.0	59.1 ± 0.0
RoBERTa	84.3 ± 0.6	81.1 ± 1.1	73.0 ± 0.0	59.8 ± 0.0
T5	84.3 ± 0.3	80.7 ± 0.8	**73.2** ± 0.1	**60.8** ± 0.8

Table 7: Encoder-only pre-trained models vs. T5's encoder in Encoder+Head classification setups.

Encoder-only Models Comparing encoder-only to encoder-decoder methods fro multi-label text classification in a fair manner is non-trivial since inherently encoder-only pre-trained models like BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) are trained on different data and with a different objective than the encoder-decoder model T5. Using T5's encoder for encoder-only methods circumvenes this problem but introduces another concern: that this encoder was trained in an encoder-decoder architecture and may thus be handicapped in comparison to encoders trained in an encoder-only architecture.

In Table 7, we present development results on UKLEX (L1) and BIOASQ (L2) for encoder-only classifiers trained from BERT, RoBERTa and T5's encoder.[11] We observe mixed results with BERT performing best on UKLEX (L1) and T5 performing best on EURLEX (L2), with absolute differences between the three models being relatively small and on average between the two datasets, favouring T5. We thus conclude that T5's encoder makes for a fair and strong encoder-only baseline and use it in subsequent experiments.

5.3 Main Results

In Table 6, we present test results for all methods trained from T5-Base.[12] The overall best performing approach is T5Enc, followed by Seq2Seq, LWAN and then Encoder+Head. The trend is thus for encoder-decoder approaches (T5Enc and Seq2Seq) to outperform encoder-only approaches (LWAN and then Encoder+Head), which use just half the model parameters. This result corroborates and considerably substantiates the observations of Liu et al. (2021). We gain further insights through a breakdown by metric and label granularity.

The advantage of encoder-decoder methods can be especially seen across macro-F1 scores, where both T5Enc and Seq2Seq outperform encoder-only approaches almost categorically (the one exception being UKLEX (L1)). This indicates that encoder-decoder approaches are particularly good at assigning less frequent labels, which is a key challenge in multi-label classification. This reading of the results is further reinforced by the observation that the performance gap increases from Level 1 datasets, which contain a smaller number of labels, to Level 2 datasets, which contain more and thus on average less frequent labels. The most striking performance gap we observe measures 7 p.p. between LWAN and Seq2Seq on MIMIC (L2).

Between the two encoder-decoder approaches, we see that the non-autoregressive use of the T5 decoder is more effective (T5Enc) than the conditional generation of labels (Seq2Seq), the gap

[11] We use the prepended `[CLS]` token representation for BERT and RoBERTa.

[12] We present development results in Table 11 in Appendix A for completeness.

between the two methods growing from Level 1 to Level 2 datasets. In the case of T5Enc, the decoder serves to build representations for all labels relevant to a dataset and in this sense defines and constraints the output space for the task. Meanwhile, in the Seq2Seq approach the model has to learn the constraints on the output space during training, and as such it is likely more prone to errors.

These main results give us a general idea of how the different approaches compare, indicating clearly that encoder-decoder approaches are superior. In subsequent sections we explore the source of performance and the limitations of encoder-decoder approaches further.

5.4 Model Capacity

One possible explanation for the stronger performance of encoder-decoder methods is that they operate with twice as many parameters as encoder-only methods. Here, we test whether this alone is the source of their improved performance, by training models from different T5 models: small, base and large.[13] Since we previously saw that trends in results are similar across L1 and L2 datasets, and more pronounced in the latter, we carry out this set of experiments on L2 datasets only. We include the stronger performing encoder-only approach, LWAN, as well as both encoder-decoder approaches. Results on the micro-F1 metric are presented in Figure 2, and on the macro-F1 metric in Figure 3 in Appendix A.[14]

Firstly, we note that T5Enc consistently outperforms the other approaches across different model sizes, in line with earlier findings (see Table 6). We also see that all methods appear to scale, with steady improvements in performance observed across increasing model sizes.

Comparing models of similar size (i.e., models with the same number of layers), we gain a more precise idea of how methods compare. Here, T5Enc still proves to be the superior approach, with T5Enc-Small outperforming LWAN-Base on 3 out of 4 datasets (UKLEX being the exception), and similarly T5Enc-Base outperforming LWAN-Large on 3 out of 4 datasets. Notice that in these comparisons, the T5Enc variants are even at a disadvantage, having the same number of layers as the LWAN variants, but lower dimensionality. Seq2Seq models, on the other hand, underperform similarly-sized LWAN models on most comparisons in terms of micro-F1, which indicates that this approach is overall less suitable for the task.[15]

5.5 Ablations on T5Enc Decoder

Here, we analyse the contribution of different aspects of the T5Enc decoder through ablations on the decoder's depth, width and self-attention.

Decoder Depth We train T5Enc models with a varying number of decoder layers. We experiments with $N \in [1, 4, 6, 12]$. In Table 8, we report results on two datasets, UKLEX (L1) and EURLEX (L2). We observe that larger depth in the decoder contributes to performance, with the full set of decoder layers (12) performing best.

Layers	UKLEX (L1)		EURLEX (L2)	
	μ-F_1	m-F_1	μ-F_1	m-F_1
N=1	84.6 ± 0.1	81.9 ± 0.1	76.6 ± 0.1	56.9 ± 0.1
N=4	84.7 ± 0.1	81.8 ± 0.1	76.9 ± 0.1	58.1 ± 1.1
N=6	**84.8 ± 0.1**	**82.2 ± 0.1**	77.0 ± 0.1	58.4 ± 1.3
N=12	**84.8 ± 0.2**	81.9 ± 0.5	**77.1 ± 0.1**	**58.8 ± 1.4**

Table 8: Development results for different numbers of decoder layers in T5Enc.

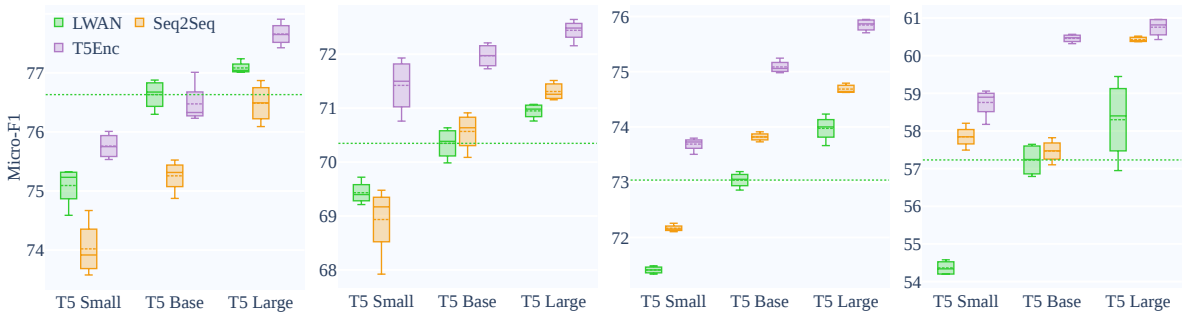
Decoder Width In this ablation, we are interested to establish the importance of label-wise representations being built in the decoder as opposed to using it to create a single output representation shared across the classification heads. To this end, we feed the decoder with a single token ID, e.g., the ID of token *'label'*, and then pass its output representation ($d \in \mathbb{R}^{dim}$) to a set of standard classification heads to produce L scores (logits), similar to the Encoder+Head method. This method can be seen as an advanced version of the Encoder+Head method that utilizes the decoder via cross-attention.

Results for Level 2 datasets are shown in Table 9 under Single-step T5Enc (Level 1 results are shown in Table 11 in the Appendix). In comparison to the Encoder+Head baseline, Single-step T5Enc is superior across the board, likely because of the added number of parameters available to the model. Compared to the standard T5Enc approach, Single-step T5Enc works slightly better for UKLEX but on all other datasets it underperforms by a large gap. We observe the same pattern for L1 results in Table 11 and thus conclude that the additional computational

[13] T5-Small has 12 layers of d=512, T5-Base has 24 layers of d=768, T5-Large has 48 layers of d=1024, where half of the layers are in the encoder and half in the decoder.

[14] All results are also presented in Table 12 in Appendix A.

[15] See Appendix A for a discussion of macro-F1 results.



(a) UKLEX (b) EURLEX (c) BIOASQ (d) MIMIC

Figure 2: Performance of the three strongest classification methods (*LWAN*, *Seq2Seq*, *T5Enc*) across three model sizes in terms of micro-F1 score. Dashed lines inside the boxes represent the mean performance across four seeds.

Method	**UKLEX (L2)**		**EURLEX (L2)**		**BIOASQ (L2)**		**MIMIC (L2)**	
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1
Encoder+Head	81.9 ± 0.6	72.9 ± 1.3	76.2 ± 0.2	54.0 ± 1.4	73.2 ± 0.1	60.8 ± 0.8	56.7 ± 0.7	22.3 ± 1.2
Single-step T5Enc	**82.6 ± 0.1**	**74.4 ± 0.8**	76.7 ± 0.2	55.8 ± 1.4	73.5 ± 0.3	61.8 ± 1.1	58.3 ± 0.5	25.8 ± 0.9
T5Enc	82.4 ± 0.4	74.2 ± 1.0	**77.1 ± 0.1**	**58.8 ± 1.4**	75.1 ± 0.0	**66.3 ± 0.1**	**60.6 ± 0.1**	31.1 ± 1.0
- No attention	81.9 ± 0.1	73.0 ± 0.5	76.8 ± 0.1	57.6 ± 0.8	74.3 ± 0.1	64.3 ± 0.3	58.6 ± 0.3	27.4 ± 1.6
- Full attention	82.3 ± 0.2	74.1 ± 0.8	**77.1 ± 0.2**	58.7 ± 0.8	**75.2 ± 0.0**	66.1 ± 0.0	**60.6 ± 0.2**	**31.6 ± 0.7**

Table 9: Ablations of T5Enc decoder. Single-step T5Enc builds a single output representation instead of L label-wise representations. Attention ablations disable (No) or fully enable (Full) the self-attention in the decoder.

power of label-wise processing is important for the good overall performance of T5Enc.

Attention Scheme The labels in multi-label classification are known to exhibit certain dependencies (Tenenboim et al., 2009; Bogatinovski et al., 2022). We measure the pair-wise dependency between labels in the four datasets included in this study, using Fisher's exact test.[16] In Table 10, we report the percentage of label pairs in Level 2 label sets for which a significant association ($p < .001$) was discovered (see Appendix A for Level 1 results). Based on the observed non-trivial rates of inter-label dependency, we hypothesize that self-attention in the T5 decoder is of key importance to the performance of T5Enc.

Level	**UKLEX**	**EURLEX**	**BIOASQ**	**MIMIC**
L2	39.5	39.7	71.2	21.3

Table 10: Percentage of Level 2 label pairs with significant association according to Fisher's exact test.

The decoder in T5 models uses *causal* attention, wherein decoder inputs can only attend to the left context. We measure the contribution of this system component by ablating it, i.e. training T5Enc models with no self-attention. In Table 9, we report results on Level 2 datasets under *No attention* (see Table 11 in Appendix A for Level 1 results). We observe that without self-attention, performance suffers considerably for all datasets, most notably so in terms of macro-F1 on MIMIC ($\Delta = 3.7$). This result indicates that self-attention indeed has a key role, although its contribution does not prove to be proportional to the rate of significant pair-wise associations in the data (Table 10)—this may be due to higher-order label dependencies taking precedence over pair-wise ones.

Having confirmed the importance of modeling label dependency above, we next consider whether we can achieve even better performance with bidirectional (rather than causal) attention in the T5 decoder. In Table 9 *Full attention*, we see that the contribution of bidirectional attention is negligible. Assuming that the model is able to adjust to the new attention scheme during the fine-tuning process, we take these results to indicate that modeling label dependency in just one direction is sufficient. Indeed, Fisher's exact test measures two-way association, disregarding the direction of the dependency.

[16]The test determines whether the observed distribution of one variable is likely to be random given the observed distribution of another variable and vice-versa.

5.6 Errors in Seq2Seq Models

The Seq2Seq approach similarly can model label dependency through self-attention and can even condition the prediction of labels on one another (in an autoregressive fashion), an ability which none of the other approaches included in this study posses. Yet, we find empirically that it underperforms T5Enc. Here, we investigate whether this finding can be explained in terms of the unconstrained output space in Seq2Seq models. Specifically, we analyse the models' predictions for the invention of novel labels.

Such errors occur for two out of the four datasets, EURLEX and UKLEX, but with extremely low frequency: the highest observed rate is 0.2% of novel labels generated for UKLEX (L2). Some examples include 'accommodation', 'domestic violence' and 'vulnerable persons'. Labels in UKLEX and EURLEX are phrased in common terms, compared to the rather technical, domain-specific labels in MIMIC and BIOASQ (see Appendix B for examples). Models trained on UKLEX and EURLEX therefore seem to interpret the output space as open-ended and on occasion generate novel labels. Still the total number of novel labels generated is negligible, so this could not explain the lower performance of this approach compared to T5Enc. The reason may instead lie with the fact that Seq2Seq models have to learn the bounds of the output space during training, whereas T5Enc models have that as a given via the fixed decoder input.

6 Conclusions

In this work, we compared four approaches to multi-label classification, two based on an encoder only and two based on an encoder-decoder. We experimented with 4 datasets from 2 different domains (legal and biomedical), which support two different label granularities. We found that encoder-decoder methods outperform encoder-only methods, in line with findings in other NLP tasks. We further found that the non-autoregressive use of an encoder-decoder model performs better than using it for conditional generation. We found that decoder depth, width and self-attention are all key contributors to the success of this best approach.

In future work, we will consider prompt-based approaches as well, specifically instruction-based fine-tuned models (Wei et al., 2022), currently limited by the excessive computational cost of encoding the full label set as part of the input string.

Limitations

Recent work has shown that models of a certain size (upwards of 3B parameters) exhibit learning properties that cannot be observed in smaller models. Due to practical limitations and environmental concerns, in this study we chose not to train models larger than T5-Large. It is thus not possible to know how emergent properties in larger models may have affected the comparison between the different approaches compared here. We believe that our findings will nevertheless be useful to NLP practitioners who operate on a constrained compute budget and may thus opt for moderately-sized models anyway.

We compare encoder-only and encoder-decoder models for multi-label classification. Decoder-only models (Radford et al., 2019) are omitted since at present there are no decoder-only methods for multi-label classification in the literature. While we could have adapted the Seq2Seq approach in our experiments to operate in a decoder-only context, we deem this unsuitable for the datasets we work with, as they contain long documents which will quickly cause problems for standard decoder-only models like GPT-2.

Domain-specific pre-trained language models exist for both the legal and biomedical domain, which outperform their generic counterparts when used for classification tasks. These models all have an encoder-only architecture, however, which renders them unsuitable for a comparison of encoder-only and encoder-decoder approaches to multi-label classification.

Our experiments consider datasets from the legal and biomedical domains first and foremost because there are publicly available datasets with hierarchical labelling in these domains, unlike others. Moreover, we believe that working in critical application domains is a worthy purpose and covering two such domains with two different datasets in each domain gives us a good view on how the examined methods are expected to work in such domains.

Ethics Statement

The legal and biomedical fields are both highly sensitive and have high impact on human life. In this work, we have ensured that the data we work with is sourced in compliance with the relevant regulations and are fully anonymized where necessary. The application of multi-label classification to this data carries no obvious risk as it can ease

the processing and categorization of documents in these domains, without having any direct impact on individuals involved in legal and medical matters.

Acknowledgments

We thank our colleagues at the CoAStaL NLP Lab and the anonymous reviewers for their feedback. This work was fully funded by the Innovation Fund Denmark (IFD).

References

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

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, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair EW Johnson, David J. Stone, Leo A. Celi, and Tom J. Pollard. 2017. MIMIC-III, a freely accessible critical care database. *Nature*.

Frederick Liu, Siamak Shakeri, Hongkun Yu, and Jing Li. 2021. Enct5: Fine-tuning T5 encoder for non-autoregressive tasks. *CoRR*, abs/2110.08426.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1101–1111.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2021. Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF2021)*. Springer, Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175.

Lena Tenenboim, Lior Rokach, and Bracha Shapira. 2009. Multi-label classification by analyzing labels dependencies. In *Proceedings of the 1st international workshop on learning from multi-label data, Bled, Slovenia*, pages 117–132.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, California, USA.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Additional Results

In Table 11, we present the development results for all models trained from T5-Base across all datasets.

In Table 12, we present detailed results for all L2 datasets for methods using T5-Small and Large. In Figure 3, which visualizes the macro-F1 results, we see that in comparisons between similarly-sized T5Enc models and LWAN models, the same trends hold here as observed in Section 5.3 for the micro-F1 metric: T5Enc is superior to LWAN as a method for multi-label classification.

Comparing Seq2Seq models to similarly-sized LWAN models, on the other hand, we see quite a different trend here compared to the micro-F1 results discussed in Section 5.3: Seq2Seq-Small outperforms LWAN-Base models on 2 out of 4 datasets (BIOASQ and MIMIC), and Seq2Seq-Base models outperform LWAN-Base models on all 4 datasets. This suggests that the Seq2Seq approach is especially suitable for the prediction of rare labels, which are better represented by the macro-F1 metric and particularly abundant in the BIOASQ and MIMIC datasets. We presume that as the only approach with access to the actual tokens comprising Level 2 label descriptors, Seq2Seq gains from lexical overlap between label descriptors and prior knowledge of the semantics of these tokens.

In Table 13, we show Fisher's exact test results for pair-wise association among labels in Level 1 label sets across all datasets. We see higher rates of pair-wise association, likely because of the smaller number of labels in each set.

B Dataset descriptors

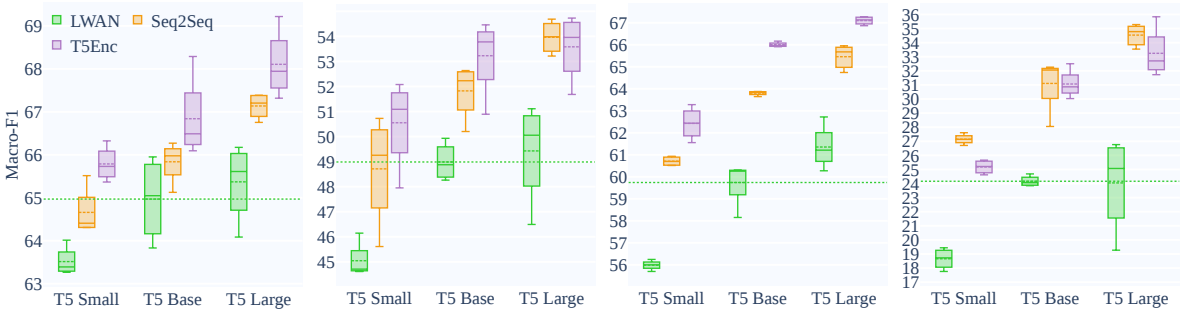
In Tables 14, 15, 16, 17, we list the original Level 1 and Level 2 label descriptors for the UKLEX, EURLEX, BIOASQ and MIMIC datasets, respectively, as well as the simplified Level 1 label descriptors, which we manually curated.

Method	UKLEX (L1)		EURLEX (L1)		BIOASQ (L1)		MIMIC (L1)	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
Encoder+Head	84.3 ± 0.3	80.7 ± 0.8	82.9 ± 0.2	72.5 ± 0.8	86.6 ± 0.0	77.1 ± 0.2	72.4 ± 0.1	65.8 ± 0.9
LWAN	84.5 ± 0.4	81.0 ± 1.1	83.0 ± 0.2	72.2 ± 0.3	86.6 ± 0.0	77.1 ± 0.3	72.5 ± 0.3	66.3 ± 1.2
Seq2Seq	84.2 ± 0.0	81.6 ± 0.2	82.8 ± 0.1	74.3 ± 0.5	86.5 ± 0.0	77.6 ± 0.2	73.2 ± 0.1	**70.3 ± 0.2**
Single-Step T5Enc	**85.1 ± 0.2**	82.4 ± 0.4	83.3 ± 0.2	73.8 ± 0.8	86.7 ± 0.1	77.1 ± 0.4	73.1 ± 0.1	67.4 ± 1.1
T5Enc	84.8 ± 0.2	81.9 ± 0.5	**83.6 ± 0.1**	**75.0 ± 0.6**	**87.0 ± 0.0**	78.1 ± 0.3	**73.6 ± 0.2**	69.2 ± 1.5
- No attention	85.0 ± 0.2	**82.5 ± 0.1**	83.5 ± 0.1	74.8 ± 0.5	**87.0 ± 0.1**	**78.3 ± 0.2**	73.6 ± 0.1	**69.5 ± 0.4**
- Full Attention	84.7 ± 0.3	82.1 ± 0.6	**83.6 ± 0.1**	**75.0 ± 0.3**	**87.0 ± 0.1**	78.0 ± 0.3	73.3 ± 0.1	68.7 ± 1.2
Method	**UKLEX (L2)**		**EURLEX (L2)**		**BIOASQ (L2)**		**MIMIC (L2)**	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
Encoder+Head	81.9 ± 0.6	72.9 ± 1.3	76.2 ± 0.2	54.0 ± 1.4	73.2 ± 0.1	60.8 ± 0.8	56.7 ± 0.7	22.3 ± 1.2
LWAN	82.0 ± 0.3	72.2 ± 0.6	76.3 ± 0.3	55.5 ± 0.8	73.2 ± 0.1	60.5 ± 0.8	57.2 ± 0.3	24.5 ± 0.4
Seq2Seq	81.2 ± 0.3	72.7 ± 1.1	75.7 ± 0.1	57.2 ± 1.1	74.1 ± 0.1	64.3 ± 0.2	57.5 ± 0.3	30.7 ± 1.7
Single-Step T5Enc	**82.6 ± 0.1**	**74.4 ± 0.8**	76.7 ± 0.2	55.8 ± 1.4	73.5 ± 0.3	61.8 ± 1.1	58.3 ± 0.5	25.8 ± 0.9
T5Enc	82.4 ± 0.4	74.2 ± 1.0	**77.1 ± 0.1**	**58.8 ± 1.4**	75.1 ± 0.0	**66.3 ± 0.1**	**60.6 ± 0.1**	31.1 ± 1.0
- No attention	81.9 ± 0.1	73.0 ± 0.5	76.8 ± 0.1	57.6 ± 0.8	74.3 ± 0.1	64.3 ± 0.3	58.6 ± 0.3	27.4 ± 1.6
- Full attention	82.3 ± 0.2	74.1 ± 0.8	**77.1 ± 0.2**	58.7 ± 0.8	**75.2 ± 0.0**	66.1 ± 0.0	**60.6 ± 0.2**	**31.6 ± 0.7**

Table 11: Development Results for all methods across datasets with T5 (base).

Method	UKLEX (L2)		EURLEX (L2)		BIOASQ (L2)		MIMIC (L2)	
	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$	μ-F$_1$	m-F$_1$
T5 (Small) models								
LWAN	75.1 ± 0.3	63.5 ± 0.3	69.4 ± 0.2	45.0 ± 0.6	71.4 ± 0.1	56.0 ± 0.2	54.4 ± 0.2	18.7 ± 0.6
Seq2Seq	74.0 ± 0.4	64.7 ± 0.5	68.9 ± 0.6	48.7 ± 1.9	72.2 ± 0.1	60.7 ± 0.2	57.8 ± 0.3	**27.1 ± 0.3**
T5Enc	**75.8 ± 0.2**	**65.8 ± 0.4**	**71.4 ± 0.4**	**50.6 ± 1.6**	**73.7 ± 0.1**	**62.4 ± 0.6**	**58.8 ± 0.3**	25.2 ± 0.4
T5 (Large) models								
LWAN	77.1 ± 0.1	65.4 ± 0.8	70.9 ± 0.1	49.4 ± 1.8	74.0 ± 0.2	61.4 ± 0.9	58.3 ± 0.9	24.0 ± 3.0
Seq2Seq	76.5 ± 0.3	67.1 ± 0.3	71.3 ± 0.1	54.1 ± 0.6	74.7 ± 0.1	65.5 ± 0.5	60.4 ± 0.1	**34.5 ± 0.7**
T5Enc	**77.7 ± 0.2**	**68.1 ± 0.7**	**72.4 ± 0.2**	53.6 ± 1.2	**75.8 ± 0.1**	**67.1 ± 0.2**	**60.8 ± 0.2**	33.2 ± 1.6

Table 12: Test Results for all methods across datasets with T5 (small) and (large).



(a) UKLEX (b) EURLEX (c) BIOASQ (d) MIMIC

Figure 3: Performance of the three strongest classification methods across three model sizes in terms of macro-F1 score. Dashed lines within the boxes represent the mean performance across four seeds.

Level	UKLEX	EURLEX	BIOASQ	MIMIC
L1	72.8	82.3	93.8	85.6

Table 13: Percentage of Level 1 label pairs with significant association according to Fisher's exact test.

Level 1 (Original)				
Agriculture and Food	Children	Criminal Law	Education	Environment
EU	Finance	Healthcare	Housing	Immigration and Citizenship
Local Government	Planning and Development	Politics	Public Order	Social Security
Taxation	Telecommunications	Transportation	-	-
Level 1 (Simplified)				
Agriculture	Children	Crime	Education	Environment
EU	Finance	Healthcare	Housing	Immigration
Local	Planning	Politics	Public	Social
Taxation	Telecom	Transport	-	-
Level 2 (Original)				
Agriculture	Air Transport	Animals	Banking	Broadcasting
Children	Citizenship	Disabled Persons	Education	Elections
Employment	Environment	EU	Finance	Fire and Rescue Services
Food	Healthcare	Housing	Immigration	Insurance
Land Registration	Local Government	NHS	Police	Pollution
Social Security	Taxation	Telecommunications	Terrorism	Urban Development

Table 14: Sample of label descriptors (Law Subject) for UKLEX dataset.

Level 1 (Original)				
Politics	European Union	International Relations	Law	Economics
Trade	Finance	Social Questions	Education & Communications	Science
Business & Competition	Environment	Transport	Working Conditions	Agriculture
Forestry & Fisheries	Agri-Foodstuffs	Production	Technology & Research	Energy
Industry	Geography	International Organisations	-	-
Level 1 (Simplified)				
Politics	International	EU	Law	Economy
Trade	Finance	Social	Education	Science
Business	Environment	Transport	Employment	Agriculture
Forestry	Food	Production	Technology	Energy
Industry	Geography	Organisations	-	-
Level 2 (Original)				
Political Framework	Political Party	Agricultural Activity	Engineering	European Organisations
Politics & Public Safety	Forestry	International Affairs	Cooperation Policy	International Security
Defence	Energy Policy	European Construction	EU Finance	Agricultural Production
Justice	International Law	Rights and Freedoms	Economic Policy	Regional Policy
Economic Structure	Trade Policy	Tariff Policy	International Trade	Marketing
Distributive Trades	Monetary Relations	Monetary Economics	Farming Systems	Food Technology

Table 15: Sample of label descriptors (EUROVOC concepts) for EURLEX dataset.

Level 1 (Original)				
Anatomy	Organisms	Diseases	Chemicals and Drugs	Analytical, Diagnostic and Therapeutic Techniques and Equipment
Psychiatry and Psychology	Phenomena and Processes	Humanities	Disciplines and Occupations	Anthropology, Education, Sociology, and Social Phenomena
Information Science	Named Groups	Health Care	Technology, Industry, and Agriculture	Publication Characteristics
Geographicals	-	-	-	-
Level 1 (Simplified)				
Anatomy	Organism	Disease	Drug	Technical
Psychology	Process	Occupation	Human	Social
Information	Groups	Healthcare	Technology	Publications
Geography	-	-	-	-
Level 2 (Original)				
Musculoskeletal System	Digestive System	Respiratory System	Urogenital System	Endocrine System
Cardiovascular System	Nervous System	Sense Organs	Embryonic Structures	Cells, Fluids and Secretions
Stomatognathic System	Hemic and Immune Systems	Tissues	Integumentary System	Plant Structures
Fungal Structures	Bacterial Structures	Viral Structures	Biomedical and Dental Materials	Microbiological Phenomena
Equipment and Supplies	Psychological Phenomena	Dentistry	Mental Disorders Behavior and Behavior Mechanisms	

Table 16: Sample of label descriptors (MeSH concepts) for BIOASQ dataset.

Level 1 (Original)				
Infection and Parasitic Diseases	Diseases of The Genitourinary System	Endocrine Nutritional and Metabolic Diseases and Immunity Disorders	Diseases of Blood and Blood Forming Organs	Mental Disorders
Diseases of Nervous System and Sense Organs	Diseases of The Circulatory System	Diseases of The Respiratory System	Diseases of The Digestive System	Neoplasms
Complications of Pregnancy, Childbirth and the Puerperium	Diseases of The Skin and Subcutaneous Tissue	Diseases of The Musculoskeletal System and Connective Tissue	Certain Conditions Originating In The Perinatal Period	Congenital Anomalies
Symptoms, Signs and Ill-Defined Conditions	Injury and Poisoning Supplementary Factors Influencing Health Status and Contact With Health Services	Supplementary Classification of External Causes of Injury and Poisoning	-	-
Level 1 (Simplified)				
Infections	Cancer	Metabolic	Blood	Mental
Nervous	Circular	Respiratory	Digestive	Urinar
Pregnancy	Skin	Muscle	Birth	Newborn
Symptoms	Injury	External	-	-
Level 2 (Original)				
Osteopathies, Chondropathies, and Acquired Musculoskeletal Deformities	Bulbus Cordis Anomalies and Anomalies of Cardiac Septal Closure	Hereditary and Degenerative Diseases of The Central Nervous System	Poliomyelitis and Other Non-Arthropod-Borne Viral Diseases of Central Nervous System	Tuberculosis
Viral Diseases Accompanied By Exanthem	Arthropod-Borne Viral Diseases	Rickettsioses and Other Arthropod-Borne Diseases	Syphilis and Other Venereal Diseases	Mycoses
Hereditary Hemolytic Anemias	Acquired Hemolytic Anemias	Aplastic Anemia and Other Bone Marrow Failure Syndromes	Other and Unspecified Anemias	Coagulation Defects
Personality Disorders, and Other Nonpsychotic Mental Disorders	Congenital Anomalies of Eye	Inflammatory Diseases of The Central Nervous System	Human Immunodeficiency Virus	Neurotic Disorders
Disorders of The Peripheral Nervous System	Disorders of The Eye and Adnexa	Diseases of The Ear and Mastoid Process	Chronic Rheumatic Heart Disease	Acute Rheumatic Fever
Ischemic Heart Disease	Diseases of Pulmonary Circulation	Acute Respiratory Infections	Chronic Obstructive Pulmonary Disease and Allied Conditions	Pneumonia and Influenza
Intestinal Infectious Diseases	Anencephalus and Similar Anomalies	Other Congenital Anomalies of Nervous System	Zoonotic Bacterial Diseases	Intellectual Disabilities

Table 17: Sample of label descriptors (ICD-9 codes) for MIMIC dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section has no number (Limitations)

☑ A2. Did you discuss any potential risks of your work?
Section has no number (Ethics statement)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5

☑ B1. Did you cite the creators of artifacts you used?
4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
To the best of our knowledge, the data is free to use for research purposes

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics statement

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Three out of 4 datasets are open access public documents that don't concern individuals. The 4th is explicitly anonymized and we trust the anonymization applied by the creatorrs

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not relevant

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our experiments are rather small-scale

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.