RAIN: Towards Application-driven Benchmarking for Natural Language Processing

Max Bartolo[†] **Jiangbo Shangguan**[‡] **Kamil Tylinski**[‡]
Walter Hernandez[‡] **Editha Nemsic**[‡] **Bartosz Kultys**[‡] **Daniel Hoadley**[‡]
Niall Roche[†‡] **Alastair Moore**[†‡] **Greg Hinch**[‡] **Pontus Stenetorp**[†]
[†]University College London [‡]Mishcon de Reya
m.bartolo@cs.ucl.ac.uk

Abstract

With increasing availability of textual data and improved model capabilities, Natural Language Processing (NLP) is gaining wider adoption in industry. However, the field is mainly guided by research-motivated benchmarks which, due to their research-oriented nature, can fail to adequately measure real-world utility of NLP applications. We envision that, in addition to existing benchmarks, more application-driven benchmarks can also help guide us towards improved Natural Language Understanding. To facilitate this, we introduce a definition of what we consider salient features for an applied benchmark and, as a first step in this direction, present the Real-World Applied Industrial NLP (RAIN) benchmark – a collection of NLP tasks and corresponding datasets with broad practical application. We formulate five new NLP tasks, collect datasets for each totalling over 150,000 annotations, and provide evaluation of baseline and task-specific models, observing a headroom gap to human performance on the overall score of 19.4%. The datasets and leaderboard are publicly available at https://rain.mdrdatascience.ai.

1 Introduction

Research in Natural Language Processing (NLP) has progressed rapidly in recent years, driven by a combination of data acquisition efforts (Bowman et al., 2015; Rajpurkar et al., 2016), advances in model architecture development (Sutskever et al., 2014; Vaswani et al., 2017), and large-scale pre-training methods (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019b).

Combined with the increased availability of textual data, this progress is contributing to NLP techniques becoming more relevant across industry. However, determining how well these systems perform is challenging since most internal datasets are not made available for research purposes. This limited access to data sourced from industry-relevant tasks restricts our ability to measure progress in terms of practical utility, and limits the domains and tasks available to the research community.[1]

Anon	**Text:** *FERC should retain its Dec 02 Day 2 start date (for LMP/financial rights congestion model...).* **Output:** *$\langle ORG \rangle$ should retain its $\langle DATE \rangle$ $\langle DATE \rangle$ $\langle DATE \rangle$ $\langle CARD \rangle$ start date (for $\langle ORG \rangle$/financial rights congestion model).*
EnergyDLT	**Text:** *Bitcoin's PoW provides, like security and maintaining the blockchain, many researchers consider finding the nonce a waste of energy being "useless".* **Output:** *$\langle BlockchainName \rangle$'s $\langle Consensus \rangle$ provides, like $\langle SecurityPrivacy \rangle$ and maintaining the $\langle Identifiers \rangle$, many researchers consider finding the nonce a $\langle ESG \rangle$ being "useless".*
JTreat	**Text:** *The relevant jurisprudence on how section 32 is to be construed is to be found in Cave v Robinson Jarvis & Rolf [2003] 1 AC 384. See especially Lord Millett at paragraph 25 to which I was referred. There is no evidence of deliberate concealment and...* **Target:** *Cave v Robinson Jarvis & Rolf* **Label:** `Positive`
JCode	**Narrative:** *Telephone call with the $\langle ORG \rangle$ regarding payment in Court.* **Label:** `JJ70 - Interim Applications`
ACode	**Narrative:** *Working on indexing all inter partes correspondence between $\langle DATE \rangle$ and $\langle DATE \rangle$* **Label:** `J09 - Plan and Review`
EmailQA	**Passage:** *Sean, The language is below. Also, Savita wanted to conference call you this afternoon to discuss this product. What time is good for you? [LBRK] Kevin [LBRK] A US Power Transaction with ...* **Q:** *Who does Savita wish to speak with?* **A:** *Sean*

Table 1: Examples from the development sets of each of the tasks in the RAIN benchmark.

Progress in NLP has traditionally been measured at a dataset-level and, more recently, evaluation benchmarks have focused on measuring general model language understanding capabilities across a range of tasks with varying complexity (McCann et al., 2018; Wang et al., 2018). In this regard, a primary consideration in the benchmark design pro-

[1] https://twitter.com/yoavgo/status/1281987787802238980

cess is dataset *difficulty* – often determined by the performance gap between humans and state-of-the-art models (Wang et al., 2019). Contemporary models perform remarkably well on these benchmarks, however, they still struggle to learn robust representations of linguistic knowledge, as evidenced by poor performance on the GLUE diagnostic set, or model susceptibility to a variety of adversarial attacks (Ettinger et al., 2017; Jia and Liang, 2017; Ebrahimi et al., 2018; Wallace et al., 2019). While adversarial approaches provide insight into the limitations of model language understanding, they still do not provide a comprehensive understanding of how models will behave in real-world applications. Application-driven datasets could pose additional challenges due to more diverse source distributions, domain adaptation requirements, naturally-occurring noise, distributional shifts over longer time-spans, and possibly more complex reasoning requirements than research-oriented datasets.

We present criteria for application-driven benchmarking and, in line with these, introduce the Real-World Applied Industrial NLP (RAIN) benchmark – a collection of five new tasks determined to have commercial application, and for which we collect data through a range of acquisition methods including expert annotation, crowdsourcing, and adversarial human annotation. The tasks are primarily sourced from the legal domain but are indicative of general applied use-cases including; general question answering on business correspondence, classifying work activities against standard coding systems, text anonymisation from various sources, and monitoring energy consumption of distributed ledger technologies. We also include one legal-specific task requiring the inference of how a case is treated in a legal judgement.

We provide empirical results for a range of baselines and observe a substantial performance gap (19.4% on the overall score) between human performance and our best baselines, suggesting an exciting direction for future research and further investigation into sourcing challenging tasks from practical use-cases. Collected datasets will be made available for research purposes, along with a public leaderboard for convenient evaluation.

In summary, our main contributions include; establishing a first comprehensive set of criteria for application-driven benchmarking, collection of five new English-language datasets for real-world tasks, totalling over 150,000 annotations, and evaluation of baseline systems demonstrating substantial headroom to human performance.

2 Related Work

NLP Benchmarks NLP progress has conventionally been evaluated at the task level, typically with large-scale, high-quality datasets being adopted as standard benchmarks. Examples include SQuAD (Rajpurkar et al., 2016, 2018) for Reading Comprehension (RC), and SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) for Natural Language Inference (NLI).

The GLUE (Wang et al., 2018) benchmark features a set of 9 tasks selected to be diverse and linguistically challenging. Progress on GLUE has been rapid with models outperforming humans on the overall score within six months of the leaderboard being made available. These performance gains, driven in part by model architecture development (Vaswani et al., 2017) and large-scale pretraining methods (Devlin et al., 2019; Liu et al., 2019a,b), motivated the development of SuperGLUE (Wang et al., 2019), specifically designed to comprise of tasks beyond existing model capabilities. Currently, the best model is just 0.5 points below human performance in terms of overall score, and humans are outdone on 3 of the 8 tasks. The BLUE benchmark (Peng et al., 2019) similarly provides a collection of five domain-specific tasks from ten corpora adopted by the BioNLP community. FLORES (Guzmán et al., 2019) extends this line of work to the construction of benchmarks for evaluating machine translation systems on low-resource languages.

Practical Application Recent performance gains and improved generalisability of NLP systems have made them increasingly more applicable to real-world problems with ongoing research contributing to progress in challenging practical applications such as online search.[2]

NLP applications are broad and range across a variety of domains such as medical (Baud et al., 1992; Murff et al., 2011; Pons et al., 2016), financial (Fisher et al., 2016), and legal (Bommarito et al., 2018; Ravichander et al., 2019; Rehm et al., 2019). Despite this increased prominence, benchmarks outside Information Extraction tend to be academic in nature, possibly due to the proprietary

[2] https://www.blog.google/products/search/search-language-understanding-bert/

Dataset	|Train|	|Dev|	|Test|	Task	Metric	Text Sources
Anon	9,348	1,965	1,991	NER (19)	F_1	Enron emails, Court Listings
EnergyDLT	3,450	349	955	Seq.Tagging (12)	F_1	Microsoft Academic Graph
JTreat	120,260	12,368	11,619	Inference (3)	Macro acc.	Case Judgements
JCode/*ACode*	33,103	2,419	3,152	Class. (13/7)	Micro acc.	Narratives (anonymised)
EmailQA	11,557	1,316	1,324	RC	F_1	Enron emails

Table 2: Datasets statistics for the five tasks included in RAIN. The number of classes is shown in parentheses.

nature of much of the work done in these domains. Recent work provides insight into the behaviour and robustness of commercial systems (Ribeiro et al., 2020; Goel et al., 2021), but does not provide a measure of performance in a real-world setting.

3 Application-driven Benchmarking

To the best of our knowledge, there is no previous definition of the requirements for an application-driven benchmark. We work with both NLP practitioners and researchers to define these criteria, and later use these as a reference for the RAIN benchmark design process.

Real-world data: Datasets should be sourced from corpora derived from or required by organic business activity. They should retain noise (such as spelling errors, abbreviations, fragmented syntax, and use of non-standard words) that naturally arises in the source texts. For example, datasets such as SQuAD (Rajpurkar et al., 2016) would not fit this condition, while datasets such as MedNLI (Romanov and Shivade, 2018) would.

Dataset size and quality: While public training data is available for a range of NLP tasks, business-specific data is less readily available. Datasets should therefore include sufficient data to train models as well as high-quality data for reliable evaluation. We recommend that each sample undergo, at minimum, distinct annotation and validation stages for expert annotation, and have at least three unique annotators if crowdsourced.

Established NLP task: Tasks should be framed within the constraints of established NLP tasks with well-defined automatic evaluation metrics. Where a task cannot satisfy this constraint, we suggest releasing it separately such that adequate effort to solve it can be made by the research community prior to its inclusion in the benchmark, allowing it to serve as a bridge between the two.

Practical application: Tasks should be relevant to everyday applications which are either carried out manually or are currently being automated in a business setting. Solving a task should be a value-adding process requiring language understanding.

Challenging: Tasks should be challenging to existing models but solvable by humans (including domain experts) – we define this as tasks on which contemporary models have not surpassed human performance. To prevent benchmark stagnation, we propose removing "solved" tasks from each subsequent iteration to provide a time-relevant snapshot of applied NLP performance.

Diverse: The benchmark should be representative of a wide range of NLP task formats, domains, languages, and applications.

License: Datasets should be licensed permissively, at minimum allowing use and redistribution for research purposes.

4 RAIN Overview

As a first step towards application-driven benchmarking guided by these criteria, and as a result of investigations into industry NLP applications and available partner domain expertise, we select a broad initial set of tasks primarily sourced from the legal domain. However, in line with the earlier *diversity* requirement, four of the five tasks represent multiple generic business-processes with cross-domain applications. The collected datasets also represent a range of NLP task structures including entity recognition, inference, multi-class classification, and reading comprehension. Examples for each task are shown in Table 1, with further examples in the appendices. Dataset statistics are summarised in Table 2.

We collect five new datasets sourced from real-world corpora from scratch, four of which are expert annotated and validated, and one which is crowdsourced using an adversarial human annotation approach (Bartolo et al., 2020). To facilitate expert annotation, we adapt the `brat` rapid annotation tool (Stenetorp et al., 2012). We set up a three-stage expert annotation data flow process;

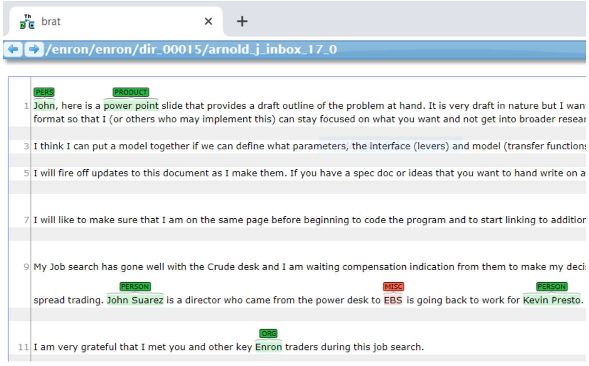

Figure 1: The annotation interface used for expert annotation of the *Anon* dataset.

First, a pre-labelling stage provides initial predictions, which are either generated from business systems or models continually re-trained as annotation efforts progress. Second, a human expert modifies labels as necessary. Domain experts are provided with initial training for task familiarity, and are then required to pass a qualification exercise. Third, a distinct expert with domain and task experience validates the work, adjusting as necessary and approving the final ground truth.

We also provide an overall benchmark metric by first averaging the *JCode* and *ACode* scores since they form part of the same task group, and then taking an equally-weighted average of the five task-group scores as the overall RAIN score.

4.1 Anonymisation (*Anon*)

Anonymisation or pseudonymisation (the identification and replacement or redaction of words) has broad application in the treatment of sensitive documents across domains (Szarvas et al., 2007; Barriere and Fouret, 2019). It presents a wide spectrum of applications that require the use and sharing of identifiable data. This is often for commercial reasons, but the General Data Protection Regulation (GDPR) has given renewed importance to redaction techniques as a means to reduce risk and assist "data processors" in fulfilling their data compliance obligations. In practice, data sharing frequently involves a tedious process of manual redaction and validation of potentially sensitive data by, for example, paralegals in the legal domain. Robust anonymisation systems can reduce the processing time required for efficient data distribution, facilitating access to de-sensitised data and encouraging data sharing.

Task Format Text anonymisation is often framed as a Named Entity Recognition (NER) task since these also represent identifiable attributes. It involves tagging input texts with the appropriate entity labels (e.g. `PERSON` or `TIME`), and poses linguistic challenges such as resolving complex co-references, or handling abbreviations and pseudonyms. We base our label set on the OntoNotes 4 annotation set (Weischedel et al., 2011) in addition to the `MISC` label from the Wikipedia corpus (Nothman et al., 2013) entity scheme to handle instances which should be redacted but do not fall within the remit of one of the other 18 entity types (see Appendix D). We measure performance using micro-averaged entity-level F_1 overlap, consistent with the evaluation of the CoNLL shared NER task (Tjong Kim Sang and De Meulder, 2003) and others (Nadeau and Sekine, 2007; Lample et al., 2016). While purely redactive systems are not necessarily concerned with exact entity type identification, this is valuable for tailored anonymisation systems as it improves model interpretability and helps increase levels of trust in the system – both important criteria for our intended application of redaction for data-sharing.

Dataset We identify two text sources which align with our criterion for *real-world data*. We source passages from the May 7, 2015 Version of the Enron email dataset (Klimt and Yang, 2004), publicly available for research purposes.[3] This repository represents one of the only substantial 'real' public email datasets. We also source two months of UK higher court listings data. These corpora include abbreviations, typographic errors, and missed punctuation and capitalisation from email correspondence and legal court listings, that we expect to reflect the text quality in similar use-cases. We pre-label these corpora using the *spaCy* NER model (Honnibal and Johnson, 2015) and regularly fine-tune a $\mathrm{BERT}_{\mathrm{Large}}$ model on the additionally collected training data to continually improve pre-labelling performance. We annotate and validate a total of 9,348, 1,965, and 1,991 labelled spans for the training, validation, and test sets using domain experts.

4.2 Energy DLT Tagging (*EnergyDLT*)

Monitoring information source changes has a range of use-cases such as market intelligence, or assessing product or technology choices. The Energy Distributed Ledger Technology Tagging (*EnergyDLT*)

[3] https://www.cs.cmu.edu/~enron/

task identifies technologies mentioned in academic publications, allowing us to assess underlying technology trends over time, with the intended application of tracking energy consumption improvements as these technologies mature.

Task Format This is a sequence tagging task where technology mentions in the text are tagged from a label set based on the second taxonomy tree level defined by Tasca and Thanabalasingham (2019). We evaluate using entity-level F_1 overlap.

Dataset The academic paper corpus is selected through a systematic literature review as set out in (Eigelshoven et al., 2020). We use the Microsoft Academic Graph (Sinha et al., 2015) for the retrieval of 43 papers and metadata in a structured format. DLT domain experts annotate and validate these, providing 3,450, 349, and 955 examples for the training, validation and test sets. Further details are provided in Appendix E.

4.3 Judicial Treatments (*JTreat*)

A reference to an external case from within a legal case report can be considered in various ways in the decision-making process. These *judicial treatments* can be characterised as a discrete set of fine-grained labels (see Appendix F), however, in most legal research applications, the core value lies in inferring whether a case is treated as *positive*, *neutral* or *negative* within the judicial report (Galgani and Hoffmann, 2010; Locke and Zuccon, 2019). Manual annotation of judicial treatments and the relationships between different legal texts allows users of data-driven commercial products to easily navigate between legal cases, find support or precedent for judgements, and perform in-depth legal research. For example, lawyers may use such treatment labels to determine whether a case is proper law, or to prioritise which decisions to examine. This task is generally carried out by human domain experts – with a considerable bottleneck being the immense effort required to retrodict or revise the label taxonomy over large and growing corpora.

Task Format The judgement data is provided by an external legal data services company and approved for public release. In that it requires three-way bi-text classification inferring how a case is treated within a judgement text, this task is similar in structure to that of Natural Language Inference (NLI). To provide a reliable evaluation of the capability to distinguish between classes that accounts for the natural class imbalance present (20% *positive*, 71% *neutral*, and 9% *negative* in the training set), we use macro-averaged accuracy as a metric.

Dataset The source data is a collection of 12,000 judgements from a commercial firm's proprietary collection of UK court judgements. Judgements are typically multiple pages long, posing an additional challenges of dealing with long form text. Treatment types (indicating the relationship between cases) are manually marked up by a team of legally trained experts. In total, we provide 120,260, 12,368, and 11,619 examples for the training, validation, and test sets.

4.4 Phase (*JCode*) & Action (*ACode*) Codes

The Jackson-codes (J-codes) set is an example of the Uniform Task Based Management System (UTBMS) codes used to classify services performed by a vendor in an electronic invoice submission. A detailed explanation is provided by Nelson and Jackson (2014).[4] *Phase* codes provide fine grained detail of the work, *Task* codes inform the type of work being carried, and *Action* codes specify how the work is done in the *Phase* and *Task*.

Time record categorisation is of particular importance in the context of digital workflows as it allows value extraction from billing data. It also facilitates effective budgeting, particularly as alternative fee arrangements become more prevalent, and improves transparency. Automatic, or machine-assisted, classification reduces administrative burden, where employees may currently each record thousands of time entries involving such codes annually. Furthermore, the adoption of UTBMS codes is often inconsistent within industries or even a given firm, with employees commonly delegating their task-based coding or assigning blocks of time entries to the same code. In these cases, there is room to improve classification quality, and allow for inter-departmental comparative analyses. There are also financial incentives as incorrect entries may be impossible to recover from counter-parties.

Task Format Both the *JCode* and *ACode* tasks involve classification of anonymised time-entry narratives (i.e., brief descriptions of the work carried out). For *JCode* there are 46 possible labels (12

[4]A similar set of codes has previously been developed in the United States. Here, the codes provide a common language for e-billing, under which both the firm and client have systems using a common code set for the delivery and analysis of bills – commonly referred to as L-codes.

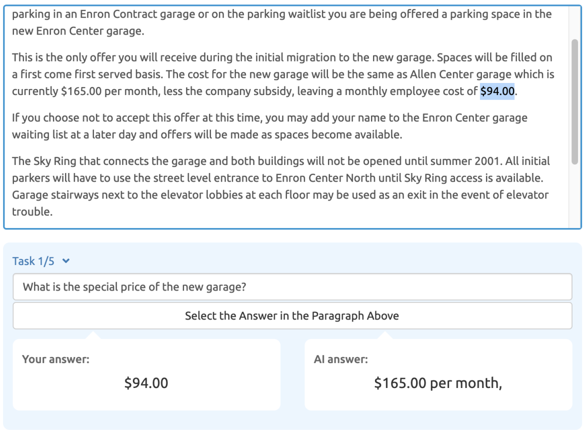

Figure 2: The interface for crowdsourcing *EmailQA*.

of which are used in the task, with an additional category that contains the rest of the labels), while there are 10 possible labels (6 of which are used in the task, with an additional category that contains the rest of the labels) for *ACode*. For detailed label definitions, see Appendices G and H. Due to their similarity, we average individual task accuracies to provide a combined score.

Dataset The source texts are time-entry narratives spanning over 500 legal matters. A narrative is typically one to two sentences providing a brief description of the work undertaken. To allow for public release, we redact entity types according to the *Anon* label set, and perform a similar three-stage pre-labelling, expert annotation and validation process. We ensure that splits respect temporal consistency by sampling sequentially in chronological order without overlap, and collect 17,305 training, 1,225 validation, and 1,626 test examples for *JCode*. *ACode* uses the same source narratives, but is slightly smaller as we remove unlabelled or ambiguous codes during validation, with 15,798 training, 1,194 validation, and 1,526 test examples.

4.5 EmailQA

Advances in machine question answering capabilities have increased the opportunity for guided or assisted support in the digital workplace. The general structure of selecting an answer to a question from a passage, such as from correspondence or documentation, has broad application across domains. Evidence-driven exploration, for example through questions such as "Did X meet Y, before event Z?", has applications in e-discovery – the act of identifying, collecting and producing electronically stored information in response to a request for production in an investigation, or for quickly locating specific events or precedents in domains such as knowledge services. RC offers applications ranging from the improved automation of business processes, to facilitating help desk operations, or supporting professional education programmes.

Task Format *EmailQA* is a RC task based on the structure of SQuAD1.1 (Rajpurkar et al., 2016) – an established benchmark. Given a passage p, in our case sourced from business-relevant emails, and a question q, the answer a is a continuous segment of text from the passage. We evaluate performance based on word-overlap F_1 score between the ground truth answer span and the model prediction – a standard RC evaluation metric. F_1 offers evaluation flexibility over Exact Match (EM) as it is tolerant to mismatch between answer and predicted spans. This is particularly desirable since we have single validated ground truth answers, rather than multiple annotations like in SQuAD. For example, for the answer *Mark Russ* and prediction *Mark*, the EM score is 0% even though the predicted answer may still be useful in practice – this is more reliably captured by the F_1 score (67% in this case).

Dataset We source passages from the Enron email corpus between 80 and 500 words and clean minor formatting issues in line with the *real-world data* requirement. We partition data splits by mailbox. Since we are interested in questions that *challenge* existing systems, we employ the adversarial human annotation approach investigated by Bartolo et al. (2020). We fine-tune $\text{RoBERTa}_{\text{Large}}$ (Liu et al., 2019b) on SQuAD1.1 using *Transformers* (Wolf et al., 2019), achieving EM and F_1 scores of $86.9\%/93.6\%$ on the SQuAD dev set, consistent with previous work. We use this model as an adversary in the annotation loop, where crowdworkers are presented with an email passage and tasked with generating and answering up to 5 questions. For each combination of passage p, question q and human annotated answer a_h, the model provides a prediction a_m which is compared against the human answer. If the model achieves an F_1 score above a threshold of 40%, the question is deemed not challenging enough to fool the model, and the process is repeated until the annotator produces a question that the model fails to answer correctly.

Crowdsourcing We use Amazon Mechanical Turk to crowdsource the data through a custom annotation interface, adapted to handle special

Model	Training Data	F_1	F_1^B
spaCy$_L$	*OntoNotes*	30.4	34.8
	Enron + CourtList	85.6	90.0
BERT$_L$	*Narratives*	40.5	53.2
	+ Enron + CourtList	**86.1**	**90.9**
	Enron + CourtList	84.7	90.0
RoBERTa$_L$	*Narratives*	42.5	52.1
	+ Enron + CourtList	85.4	89.3

Table 3: Results for models trained on different datasets, evaluated on the *Anon* test set. F_1 is entity level. F_1^B is the binary redacted vs not redacted case.

Model	Training Data	EM	F_1
	EmailQA	23.6	28.8
BiDAF	*+ SQuAD*	27.9	34.6
	+ SQuAD + $\mathcal{D}_{\text{ADV}}$	31.4	37.8
	EmailQA	42.1	49.4
BERT$_L$	*+ SQuAD*	42.7	50.3
	+ SQuAD + $\mathcal{D}_{\text{ADV}}$	50.6	57.9
	EmailQA	49.0	57.6
RoBERTa$_L$	*+ SQuAD*	50.4	58.2
	+ SQuAD + $\mathcal{D}_{\text{ADV}}$	**55.7**	**62.9**

Table 4: *EmailQA* test set results using different training data. $\mathcal{D}_{\text{ADV}}$ is the adversarially created $\mathcal{D}_{\text{BiDAF}}$, $\mathcal{D}_{\text{BERT}}$ and $\mathcal{D}_{\text{RoBERTa}}$ from Bartolo et al. (2020)

line-break tokens introduced during preprocessing. Crowdworkers are geographically restricted, must have a Human Intelligence Task (HIT) Approval Rate greater than 98%, and have successfully completed at least 1,000 HITs. We pay $2 for every HIT, collecting up to 5 questions which beat the model, with an average completion time of 876s.

Quality Control Crowdworkers are provided with an initial training and qualification task. Successful candidates proceed to work on the actual task, for which a proportion of each worker's questions are manually reviewed and validated. For the validation and test sets, we additionally require questions to be answered correctly by at least one of three further validators. We obtain answerability rates of 87.9% and 85.6% on these splits, filtering out any examples where there is no additional validator answer matching the original – this ensures that these questions are challenging RC models, while also being human-answerable. We collect 11,557 train, 1,316 validation, and 1,324 test questions at a total cost of approximately $10,000.

5 Experiments

In this section we present experiments using baseline models on the RAIN benchmark.

5.1 General Baselines

Simple Baselines We include three simple baselines; i) *Random* – which predicts a uniformly sampled class for every instance, ii) *Most Frequent* – which predicts the most frequent class, and for *EmailQA* locates the most frequent answer start span as a proportion of passage length, and answer length, and iii) *CBoW* – logistic regression on the mean of the 300D GloVe (Pennington et al., 2014) input sequence embeddings (see Appendix C).

Pre-trained LMs We fine-tune pre-trained masked language models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) independently for each task. We use cased models for *Anon* and *EnergyDLT*, and uncased models otherwise.

Domain Adaptation While only one of the five RAIN tasks relies on strictly legal source documents, we also fine-tune LegalBERT (Chalkidis et al., 2020) as a domain-adaptation baseline.

5.2 Task-specific Baselines

While our primary interest lies in evaluating general model capabilities across tasks, we also encourage submissions for models tailored to individual tasks to the RAIN leaderboard.

Anon We evaluate several task-specific models, in particular *spaCy* NER, which was used as the initial pre-labeller. Results are shown in Table 3 where we also compare against those achieved by training on the internal un-redacted *Narratives* dataset.

EmailQA We evaluate the RoBERTa$_{\text{Large}}$ model used as an adversary-in-the-loop and achieve EM/F_1 scores of 0.0%/3.8% and 0.0%/4.1% on the *EmailQA* validation and test sets, as expected by definition of the annotation setup. We further train RoBERTa$_{\text{Large}}$ on three dataset combinations; i) the *EmailQA* training set (11,557 examples), ii) *EmailQA* combined and shuffled with SQuAD1.1 (99,156 examples), and iii) *EmailQA* and SQuAD1.1 combined with additional adversarially-collected datasets $\mathcal{D}_{\text{BiDAF}}$, $\mathcal{D}_{\text{BERT}}$ and $\mathcal{D}_{\text{RoBERTa}}$ (Bartolo et al., 2020) (129,156 examples). We also compare results for BiDAF (Seo et al., 2017), and BERT$_{\text{Large}}$.

Model Metric	*Anon* F_1	*EnergyDLT* F_1	*JTreat* Macro Acc.	*JCode*/ *ACode* Acc.	*EmailQA* F_1	Overall
Random	0.3	8.2	33.2	2.3 / 11.0	2.4	10.2
Most Frequent	33.3	17.3	33.7	32.2 / 41.4	4.1	25.0
CBoW	26.0	15.5	36.0	31.4 / 69.8	4.8	26.6
BERT	**85.6**	25.2	43.0	32.7 / 75.1	57.9	53.1
RoBERTa	84.7	20.5	44.6	**34.2 / 78.9**	**62.9**	**53.9**
LegalBERT	85.3	**28.7**	**45.2**	32.2 / 74.7	44.6	51.5
Human (est.)	$90.9_{(5.3)}$	$30.8_{(2.1)}$	$65.0_{(19.8)}$	70.7 / $96.2_{(26.9)}$	$96.1_{(33.2)}$	$73.3_{(19.4)}$

Table 5: Baseline performance on the RAIN test sets. In all cases, higher is better and values are in $[0, 100]$. Numbers in parentheses show gap between best baseline and human performance.

5.3 Human Performance

We estimate human performance through an additional annotation round of expert annotation. For *JTreat*, we provide a legal expert with 50 legal citations and short case snippets, making this a conservative estimate. For *EmailQA*, we assess non-expert performance by comparing a randomly selected validator answer to the ground truth for each question. We obtain EM/F_1 scores of 76.1%/83.9% on the validation set and 68.8%/80.9% on the test set. We also manually answer 150 questions to estimate expert performance, with EM/F_1 scores of 89.3%/96.1%. This performance gap is in part explained by the passage lengths, time-constrained crowdsourcing, and expert annotator task familiarity.

6 Results and Discussion

Results on the RAIN test set are shown in Table 5. The simple baselines perform poorly across most tasks, with the exception of CBoW on *ACode* which outperforms majority class by 28.4%. BERT and RoBERTa demonstrate considerable performance gains on *Anon*, *ACode*, and *EmailQA*. For *JTreat*, both BERT and RoBERTa demonstrate an ability to distinguish between judgement types.

Despite being similar tasks requiring classification of the same anonymised narratives, we find that the *JCode* task is substantially more challenging for both models and humans, likely a result of the finer class granularity. On *EmailQA*, we find that supplementing the training data with additional adversarially sourced questions boosts performance for BERT by 7.6% F_1 and 4.7% for RoBERTa, although adding SQuAD training data only adds 0.9% F_1 for BERT, despite it being considerably larger in size (see Table 4). Our best baselines still lag behind human performance on all tasks, with an overall score difference of 19.4%. The smallest headroom gap is 2.1% on *EnergyDLT* where human performance is a conservative non-expert estimate, followed by 5.3% on *Anon*, although further experiments with the unredacted portion of the *Narratives* corpus indicate that there exist source texts for which models find anonymisation considerably more challenging.

LegalBERT shows improved performance on *JTreat* as expected due to the additional pre-training on case law documents. Surprisingly, LegalBERT also outperforms on *EnergyDLT* although it does substantially worse on *EmailQA* and underperforms BERT on the two other tasks, suggesting that solving RAIN requires overcoming challenges beyond domain adaptation.

7 Conclusion

In this work, we establish an initial set of criteria for application-driven NLP benchmarking. In line with these, we motivate five diverse new tasks, collecting datasets for each – forming the first iteration of the VALUE benchmark. We evaluate a range of baseline models and find a considerable gap to human performance of 19.4% overall suggesting that application-driven tasks can be challenging to contemporary models and interesting to the research community. We plan to continue expanding in this direction with new tasks, and the involvement of more stakeholders with applied NLP use-cases which could, for example, take the form of a workshop series. It is our hope that this work will contribute towards deepening the ongoing collaboration between industry and the NLP research community, and encourage further development and release of application-driven NLP datasets.

Ethics Considerations

The source data for *Anon*, *EnergyDLT*, *JTreat*, and *EmailQA* is publicly available English text. *JCode*/*ACode* is based on English text written by legal professionals in their day-to-day work. While the text is a product of the legal professional-client relationship, they are a part of the billing process and concerns work actions. As these texts have been automatically and then manually anonymised as per Section 4.3 to ensure the privacy of both the legal professional and client.

Our datasets are motivated by the work process, rather than specific cases, of an organisation employing NLP-based tools. As such, we have good reason to believe that the impact of the datasets we introduced will not have a negative effect on vulnerable populations as their characteristics and specifics are not contained within the data.

References

W Brian Arthur. 2007. The structure of invention. *Research Policy*, 36:274–287.

Valentin Barriere and Amaury Fouret. 2019. May I Check Again? A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. Technical report.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotations for Reading Comprehension. *arXiv preprint arXiv:2002.00293*.

RH Baud, A-M Rassinoux, and J-R Scherrer. 1992. Natural language processing and semantical representation of medical texts. *Methods of information in medicine*, 31(02):117–125.

II Bommarito, J Michael, Daniel Martin Katz, and Eric M Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *arXiv preprint arXiv:1806.03688*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

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.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Felix Eigelshoven, André Ullrich, and Benedict Bender. 2020. Public blockchain - a systematic literature review on the sustainability of consensus algorithms. In *ECIS*.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*.

Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214.

Filippo Galgani and Achim Hoffmann. 2010. Lexa: Towards automatic legal citation classification. In *Australasian Joint Conference on Artificial Intelligence*, pages 445–454. Springer.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel Locke and Guido Zuccon. 2019. Towards automatically classifying case law citation treatment using neural networks. In *Proceedings of the 24th Australasian Document Computing Symposium*, ADCS '19, New York, NY, USA. Association for Computing Machinery.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Harvey J Murff, Fern FitzHenry, Michael E Matheny, Nancy Gentry, Kristen L Kotter, Kimberly Crimin, Robert S Dittus, Amy K Rosen, Peter L Elkin, Steven H Brown, et al. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, 306(8):848–855.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26.

David Nelson and Jackson. 2014. EW-UTBMS Civil Litigation J-Code Set Overview and Guidelines.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, 194:151–175.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

M. E. Porter and V. E. Millar. 1985. How information gives you competitive advantage. *Harvard Business Review*, 64(4):149–160.

R. Pujadas, M. Thompson, W. Venters, and S. Wardley. 2019. Building situational awareness in the age of service ecosystems. *Harvard Business Review*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Georg Rehm, Julián Moreno-Schneider, Jorge Gracia, Artem Revenko, Victor Mireles, Maria Khvalchik, Ilan Kernerman, Andis Lagzdins, Marcis Pinnis, Artus Vasilevskis, Elena Leitner, Jan Milde, and Pia Weißenhorn. 2019. Developing and orchestrating a portfolio of natural legal language processing and document curation services. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 55–66, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *The International Conference on Learning Representations (ICLR)*.

A. Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, B. Hsu, and K. Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW '15 Companion*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. 2007. State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association*, 14(5):574–580.

Paolo Tasca and Thayabaran Thanabalasingham. 2019. A taxonomy of blockchain technologies: Principles of identification and classification. *Ledger*, 4.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

A VALUE Chains

With the proliferation of models for NLP tasks, it is harder to understand the performance differences between models and we believe benchmark composition will play an important role in practical applications. Here we discuss the motivation for the composition and dependancies of RAIN benchmark tasks.

We consider a business to be a purposed system, where each ML process links some purpose or market need with a data dependant effect that can be exploited to satisfy it (Arthur, 2007). This idea of a *value chain* is based on the process view of organizations, made up of components each with inputs, transformation processes and outputs (Porter and Millar, 1985). We treat datasets, tasks and engines as fundamental components of the system. Hence the benchmark is driven by the "NLP needs" of a company and captures the relationships between tasks on different categorise of activities within the business value chain.

Figure 3 shows the distribution of selected tasks within a business value chain as a Wardley Map (Pujadas et al., 2019). The RAIN benchmark is not only diverse in NLP task structures (see Table 2) but also represents a diverse set of tasks across a value chain.

EmailQA and *Anon* represent tasks within the marketing and operational functions of a firm. *EnergyDLT* and *JTreat* are domain-specific tasks that relate to knowledge management. *JCode* and *ACode* represent tasks within the finance functions of a firm relating to budgeting or management accounting.

Figure 4 shows the dependencies between components of the system. We make several observations: First, we consider both endogenous and exogenous datasets. Secondly, some tasks are dependant on multiple datasets e.g. *Anon* is dependant on CourtList and Enron. Third, some tasks are dependent on the engine output from upstream tasks e.g. *JCode* is dependant on the *Anon* engine. The *Anon* engine is shown at two stages of maturity, *Anon*[0] and *Anon*[1], before and after supervised learning. Fourth, datasets that originate in one area of the business can be used to support tasks in other functional areas e.g., the *Narratives* dataset, supports the *JCode* task.

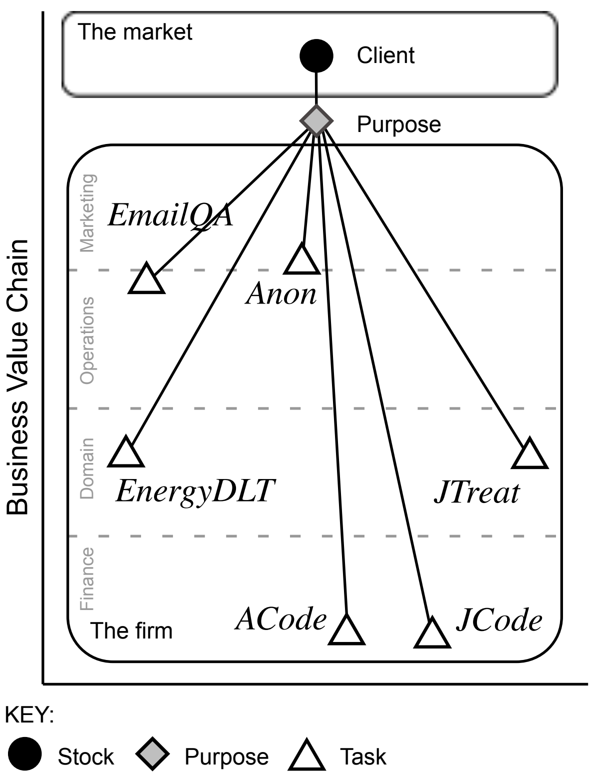

Figure 3: Business value chains. Distribution of selected NLP tasks across a value chain ordered by business function. Each individual task contributes to an overall business purpose.

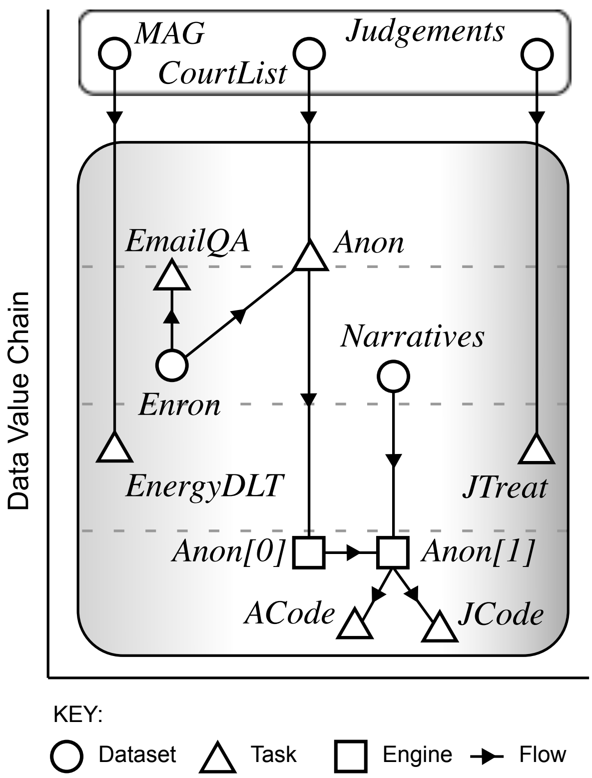

Figure 4: Data value chains. Showing the dependencies between datasets, tasks and engines in the context of a business system. We consider both endogenous and exogenous datasets.

B Progress on NLP benchmarks

We motivate exploring new benchmarks in part by highlighting the rapid recent progress on multi-task benchmarks.

The GLUE (Wang et al., 2018) benchmark features a set of nine tasks including single sentence, inference, and similarity and paraphrase tasks, selected to be diverse and linguistically challenging. Progress on GLUE has been rapid with state-of-the-art models outperforming humans on the overall score within a six month period, see Figure 5. Models currently outperform humans on 5 of the 9 tasks, and the largest performance gap between machines and humans is just 1.4% accuracy on the Winograd NLI task. Progress on SuperGLUE (Wang et al., 2019) has been similarly impressive, with the best models recently outperforming human baselines.

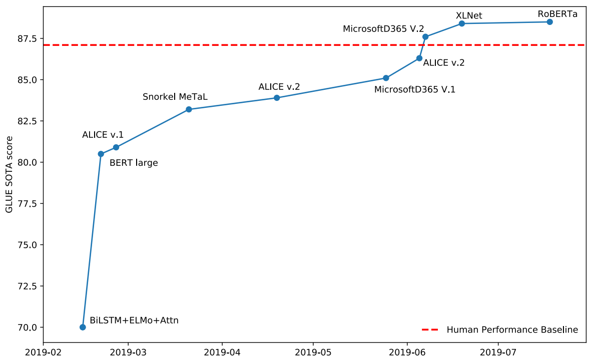

Figure 5: Progress on the GLUE benchmark. Rapid improvement over a six-month period where state-of-the-art models approach and surpass human performance.

The RAIN benchmark provides a diverse number of tasks across the value chain (see Figure 3) and will develop a representative task dependency graph across a range of business functions.

This will allow researchers to better explore the relationship between datasets, tasks and models. Current research-led benchmarks struggle to accurately capture the intricacies of their intended application domain due to various challenges such as access to data. Open research questions related to application-driven benchmarks include the relationship of tasks in a multi-task learning setting, and efficient parameter transfer across task sets or structural dependencies of upstream and downstream engines. Regardless of these future directions, the gap to human performance on RAIN gap suggests that application-driven tasks can be challenging to contemporary models and of considerable interest to the research community.

C CBOW Baseline Details

We use mean 300D/840B GloVe embeddings as sequence representations. We concatenate both input sequences for bi-text classification. For sequence tagging we classify each token independently and aggregate. For Reading Comprehension we classify the start and end tokens separately, which we find works better than classifying the start token and answer length.

D *Anon* Task

We base our annotation scheme on OntoNotes 4 (Weischedel et al., 2011), but we include the `MISC` label from the entity scheme based on the Wikipedia corpus (Nothman et al., 2013) to handle instances which should be redacted but do not fall within the remit of one of the other 18 entity types.

For example, the `MISC` label could be used where the word is either ambiguous in a given context (e.g. could be `PERSON` or `ORG`) or do not fall under another type but require redaction (e.g. a telephone number or case code).

Evaluation of human performance on the Anonymisation task and Action codes classification have been assessed by comparing the labels and classes provided by the first annotator with the ground truth, which have been validated by a second, more senior expert annotator.

Human performance for Action code classification is 96.7% on validation set and 96.2% on test set. For anonymisation, estimated human performance (F_1 score) is 98.7% on the validation set and 90.9% on the test set.

E *EnergyDLT* Task

The annotation pipeline is divided into four main processes. Each process feeds inputs to sub-tasks or sub-pipelines at different points; i) The metadata acquisition process uses Microsoft Academic Graph (Sinha et al., 2015) for the retrieval of paper metadata in a structured format; ii) The data acquisition process retrieves the selected papers; and iii) The data processing handles the generation of files to use for annotation. We use AWS services for the storage of the processed data at different stages of the annotation pipeline and the execution of sub-pipelines for training and pre-labelling.

We have relied on the distributed ledger technology taxonomy proposed by (Tasca and Thanabalasingham, 2019) when structuring the task and the label set. The dataset has a tree structure, there are 136 labels, aggregated into 12 groups. We are only using the top level group of labels in the dataset.

The additional group of labels that we have created relate to `Identifiers`; information such as other names of a blockchain, date of creation, token names, creator information, and purpose. Another additional group of labels that we have created is `MISC`, similarly as in *Anon* case, this group is meant to capture any entities that are relevant to the DLT topic and is not covered by the other groups.

F *JTreat* Task

Judicial treatments can be characterised as a discrete set of labels over the types of relationship between legal cases.

Human performance evaluation of judgements treatment have been performed by providing the legal practitioner with 50 judgements and them assessing the treatment of the last citation in the judgement. Based on the high level of care taken in annotating the data by legally-trained experts, and the elimination of any examples considered ambiguous, we assume that subject matter experts should be able to identify the treatment of all examples used in the dataset perfectly if they were not constrained by time restrictions.

The judgement treatment classes represents a more granular version of the NLI task, hence relate to course labels: such as `Positive`, `Negative`, and `Neutral`. `Positive` classes are `APPLIED`, `AFFIRMED`, `APPROVED`, `FOLLOWED`, whereas `Neutral` are `CITED`, `CONSIDERED`, `REFERRED TO`, `RELIED UPON`. `Negative` category consists of the following classes: `DISAPPROVED`, `DISTINGUISHED`, `NOT APPLIED`, `NOT FOLLOWED`, `OVERRULED`, `REVERSED`.

G *JCode* Task

The source narratives have been collected from the time recording system of a partner law firm, anonymised as previously described, and manually reviewed.

Effect of Sampling Strategy We also explored the effect of sampling strategy during dataset construction. We compare stratified sampling of the label set with temporal sampling. Temporal sampling ensures that data splits respect chronological order such that all instances in the validation set are generated further forward in time than those in the training set, and those in the test set being further than those in the other two splits. Results for this experiment can be seen in Table 6.

Model	Sampling Type	Dev acc.	Test acc.
$BERT_B$	Stratified	51.3	52.3
	Temporal	34.5	32.8

Table 6: Results for different partitions of Narratives for the *Anon* task constructed with different sampling methods on the 20,844 Train/dev sets. Temporal sampling results in lower performance, but is more realistic of deployed scenarios and is chosen for the final RAIN benchmark.

Additional dev set examples for the *JCode* task are shown in Table 7

The complete list of *JCode* labels and their definitions is seen in Table 8. For the purpose of this task we use a subset from all available labels, see Table 8 for details.

JCode	**Narrative:** *Attendance on- Meeting $\langle MISC\rangle$ and $\langle MISC\rangle$. Calls $\langle MISC\rangle$. Corr. Re $\langle ORG\rangle$.* **Label:** `JC10 - Pre-Action Factual investigation`
JCode	**Narrative:** *Discussions with $\langle MISC\rangle$ re letter of opinion; Researching $\langle LAW\rangle$ and $\langle LAW\rangle$; drafting letter of opinion.* **Label:** `JC20 - Pre-Action Legal investigation`
JCode	**Narrative:** *Drafting email to $\langle PERS\rangle$ re $\langle NORP\rangle$ licence.* **Label:** `JF10 - Disclosure Preparation of Disclosure Report/Disclosure Proposal`
JCode	**Narrative:** *Telephone call with the $\langle ORG\rangle$ regarding payment in Court.* **Label:** `JJ70 - Interim Applications`
JCode	**Narrative:** *Meeting with $\langle PERS\rangle$ to discuss amendments to the $\langle ORG\rangle$ representation agreement.* **Label:** `JP10 - Out of Scope Work Outside Lit. Procedural Stages`

Table 7: Additional examples from the *JCode* dev set.

Class	Phase and Task Description
`JC10`⋆	PRE-ACTION Factual Investigation
`JC20`⋆	PRE-ACTION Legal Investigation
`JC30`⋆	PRE-ACTION Pre-action Protocol or Similar Work
`JC40`	PRE-ACTION Group Litigation Book Building
`JA10`	FUNDING Funding
`JE10`⋆	STATEMENTS OF CASE Issue Serve Proceedings and Preparation of Statement of Case

Continued on next column

Continued from previous column

Class	Phase and Task Description
JE20	STATEMENTS OF CASE Review of Other Partys Statement of Case
JE30	STATEMENTS OF CASE Requests for Further Information
JE40	STATEMENTS OF CASE Amendment of Statements of Case
JB10	BUDGETING COSTS ESTIMATE Budgeting Own Sides Costs
JB20	BUDGETING COSTS ESTIMATE Precedent H
JB30	BUDGETING COSTS ESTIMATE Budgeting Between the Parties
JB40	BUDGETING COSTS ESTIMATE Monitoring Cost Budgets
JI10	CMC Case Management Conference
JI30	COSTS MANAGEMENT HEARING Costs Management Hearing
JF10⋆	DISCLOSURE Preparation of Disclosure Report Disclosure Proposal
JF20	DISCLOSURE Obtaining and Reviewing Documents
JF30	DISCLOSURE Preparing Serving Disclosure Lists
JF40	DISCLOSURE Review of Other Sides Disclosure
JG10⋆	WITNESS STATEMENTS Preparing Witness Statements
JG20	WITNESS STATEMENTS Reviewing Other Partys Witness Statements
JH10	EXPERT REPORTS Own Expert Evidence
JH20	EXPERT REPORTS Other Party's Expert Evidence
JH30	EXPERT REPORTS Joint Expert Evidence
JI20	PTR Pre-Trial Review
JK10	TRIAL PREPARATION Preparation of Trial Bundles
JK20⋆	TRIAL PREPARATION General Preparation for Trial
JL10	TRIAL Advocacy
JL20	TRIAL Support of Advocates
JL30	TRIAL Judgement and Post-Trial
JD10	ADR SETTLEMENT Mediation
JD20⋆	ADR SETTLEMENT Other Settlement Matters
JJ10	INTERIM APPLICATIONS Originating Process or Statement of Case or Default or Summary Judgement
JJ20⋆	INTERIM APPLICATIONS Injunction or Committal
JJ30	INTERIM APPLICATIONS Disclosure or Further Information
JJ40	INTERIM APPLICATIONS Evidence
JJ50	INTERIM APPLICATIONS Costs Only
JJ60	INTERIM APPLICATIONS Permission Applications
JJ70⋆	INTERIM APPLICATIONS All Other Applications Not Covered Above
JM10	COSTS ASSESSMENT Preparing Costs Claim
JM20	COSTS ASSESSMENT Points of Dispute Replies and Negotiations
JM30	COSTS ASSESSMENT Hearings
JM40	COSTS ASSESSMENT Post Assessment Work Excluding hearings

Continued on next column

Continued from previous column

Class	Phase and Task Description
JN10	OUT OF SCOPE WORK Outside Scope Agreed with Client
JO10	OUT OF SCOPE WORK Outside Court Approved Budget
JP10⋆	OUT OF SCOPE WORK Outside Litigation Procedural Stages
OTHER⋆	Other J-Codes

Table 8: List of *JCode* classes with a total of 46 labels. The asterisk indicates the labels that were used for the task. `Other` consists of the remaining labels - the ones which are not marked with asterisk.

H *ACode* Task

The lowest tier of the Jackson code taxonomy is the `Action` code. `Actions` specify how the work is done in the previous two tiers.

The complete list of *ACode* labels and their definitions is presented in Table 9. For the purpose of this task we use a subset from all available labels, see Table 9 for details. The dataset has been collected from the time recording system of the law firm and manually reviewed.

Class	Description
J01⋆	Client Communications
J02⋆	Counsel Communications
J03⋆	Other Side Communications
J04	Witness Communications
J05	Expert Communications
J06⋆	Internal Communications
J07⋆	Other External Communications
J08	Appear For Attend
J09⋆	Plan Prepare Draft Review
J10	Billable Travel Time
OTHER⋆	Other Action Codes

Table 9: List of *ACode* classes with a total of 10 labels. The asterisk indicates the labels that were used for the task. `Other` consists of the remaining labels - the ones which are not marked with asterisk.

Additional dev set examples for the *ACode* task are shown in Table 10.

ACode	**Narrative:** *Attendance on client - ⟨PERS⟩ and ⟨PERS⟩, with ⟨MISC⟩ for conference call.* **Label:** `J01 - Client Communications`
ACode	**Narrative:** *Attendance on ⟨ORG⟩ - review of letter received and considering Order made. Attendance on ⟨PERS⟩ re. the same.* **Label:** `J03 - Other Side Communications`
ACode	**Narrative:** *Attendance on ⟨PERS⟩ with instructions to update counsel with ⟨PERS⟩ report.* **Label:** `J06 - Internal Communications`
ACode	**Narrative:** *Attendance on Counsel - review of correspondence from ⟨ORG⟩'s clerk and instructions to ⟨PERS⟩ re. the same.* **Label:** `J02 - Counsel Communications`
ACode	**Narrative:** *Preparing letter dated ⟨DATE⟩ and bundle amendments with ⟨PERS⟩ for Court.* **Label:** `J09 - Plan Prepare Draft Review`

Table 10: Additional examples from the *ACode* dev set.

I *EmailQA* Task

One of the limitations of *EmailQA* is that the data was not annotated by experts who would make day-to-day use of such systems, therefore, despite the passages being sourced from a highly-relevant domain, the questions are not guaranteed to represent the types of questions which would naturally be asked of an in-production question answering system.

Additional dev set examples for the *EmailQA* task are shown in Table 11.

EmailQA	**Passage:** *. . . an options trader wants to be long vol outside the trading range, believing that a breakout of the range leads to volatility while trying to find new equilibrium. supports a vol smile theory. in addition, in some commodities realized vol is a function of price level. nat gas historically is more volatile at $5 than at $4 and more volatile at $4 than $3. thus there has been a tendency for all calls to have positive skew and all puts except . . .* **Question:** *Is natural gas less volatile at $3 or $5?* **Answer:** *$3*
EmailQA	**Passage:** *. . . has signed on to host a wine tasting dinner at Cafe Brand of Jersey City this Wednesday October 24 at 6:30PM. The $70 entry fee includes tax, gratuity and an elegant "World Bistro" style 7-course meal prepared by Executive Chef Seth Coburn . . .* **Question:** *The event is being held in which month?* **Answer:** *October*
EmailQA	**Passage:** *Jeff, [LBRK] Thanks for the update. [LBRK] Kevin and I are on point to provide input for Origination to the proposed Wednesday filing. Please be sure . . . and how and when will Enron's position be submitted to the CPUC/Utilities? [LBRK] Regards, [LBRK] Lamar* **Question:** *Who is currently ready with Kevin for the proposed Wednesday filing?* **Answer:** *Lamar*
EmailQA	**Passage:** *. . . for CAISO imbalance energy. [LBRK] My comments are: [LBRK] The FERC is able to only order refunds to jurisdictional entities and, given appeals, it may take years before the full extent of thre refunds are known. Therefore there will be a significant difference between the change in the mitigated market price (MMP) as declared by FERC and the PX credit . . .* **Question:** *What issue would the company have when trying to get some of their money back?* **Answer:** *The FERC is able to only order refunds to jurisdictional entities*

Table 11: Examples from the *EmailQA* dev set.