Auto-Citations

By: Arian Houshmand and CJ DuHamel

What is the big idea?

We want to add in text citations to a paper based upon a list of papers we used to write the paper

The Data

And how we constructed our dataset

Original Datasets

S2ORC (The Semantic Scholar Open Research Corpus)

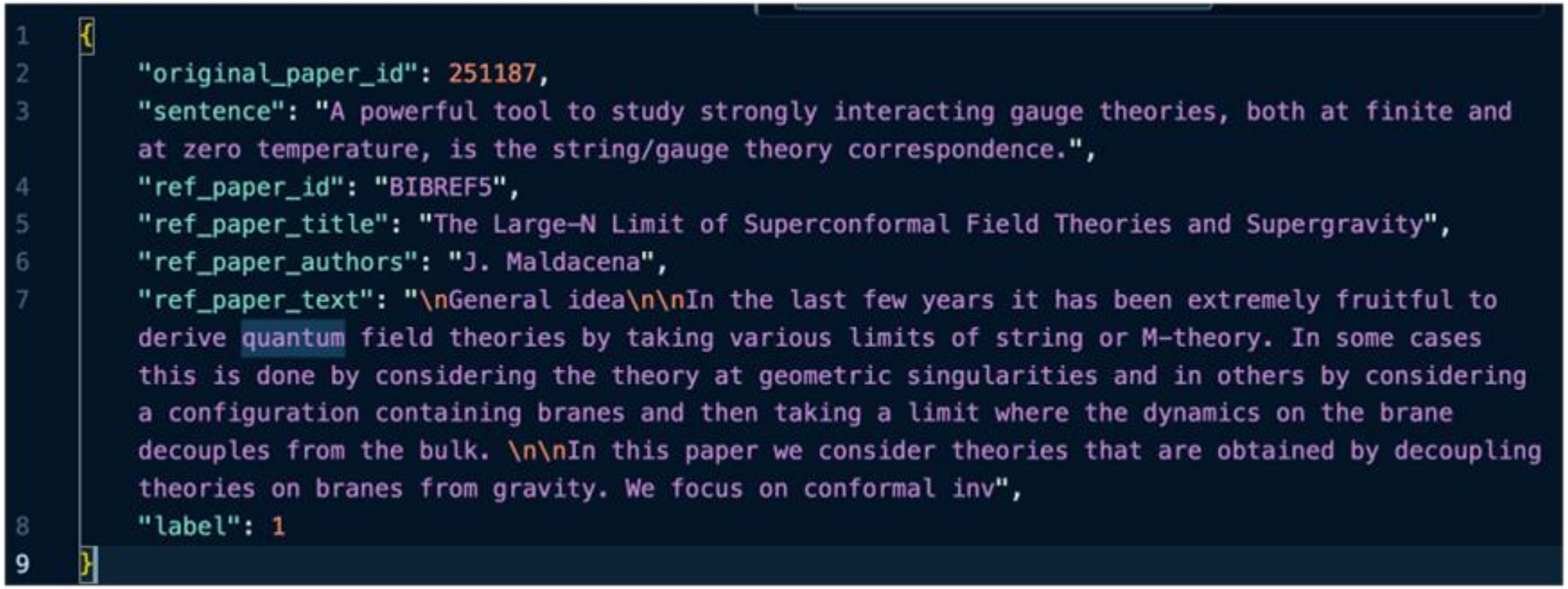
- 500+ Gb of Compressed JSONL Files
- Each Line Corresponds to a paper
- Papers are parsed into structured segments such as body text, bibliography, etc.
  - However, not all papers are structured the same way
- Annotations (sometimes) indicate the byte locations of figures, citations, and other notes

CiteWorth

- Derived from the S2ORC
- Each row corresponds to a paragraph in an original paper
- Each sentence is tagged with:
  - A label indicating “cite-worthiness”
  - the bibliography references for the papers that were cited

Building the Dataset – The S2ORC Dataset

- The S2ORC dataset is split into 1 Gb shards, downloading took ~26 Hours due to throttling
  - Signed Keys expired before completion!
- To make the search for papers easier we build a mapping from paper_id to shard_filename
  - This also allows us to quickly determine if a paper exists in our dataset or not



```
{
    "original_paper_id": 251187,
    "sentence": "A powerful tool to study strongly interacting gauge theories, both at finite and at zero temperature, is the string/gauge theory correspondence.",
    "ref_paper_id": "BIBREF5",
    "ref_paper_title": "The Large-N Limit of Superconformal Field Theories and Supergravity",
    "ref_paper_authors": "J. Maldacena",
    "ref_paper_text": "\nGeneral idea\n\nIn the last few years it has been extremely fruitful to derive quantum field theories by taking various limits of string or M-theory. In some cases this is done by considering the theory at geometric singularities and in others by considering a configuration containing branes and then taking a limit where the dynamics on the brane decouples from the bulk. \n\nIn this paper we consider theories that are obtained by decoupling theories on branes from gravity. We focus on conformal inv",
    "label": 1
}
```

Building the Dataset – Integrating CiteWorth

Dataset Info

- Class Distribution
  - Label 0 (Reference not cited): 68.4%
  - Label 1 (Reference is cited): 31.6%
- Missing Values
  - Papers Missing Title: 48.86%
  - Papers Missing Authors: 48.86%
  - Papers Missing Text: 0.1%

The Model

Designing and Implementing the Auto-Citation Machine

Text Representation

- We structure the input as follows:

[CLS] *original_sentence* [SEP] *ref_title* *ref_authors* *ref_text* [SEP]

Model

- LR + TFID
  - Val : acc = 52%, prec = 32%, rec = 46%, f1 = 38%
  - Test = acc = 50%, prec = 30%, rec = 43%, f1 = 35.8
- Sci-Bert (base model)
  - Val : acc = 64%, prec = 50%, rec = 50%, f1 = 50%

Threshold Tuning + Grid Search

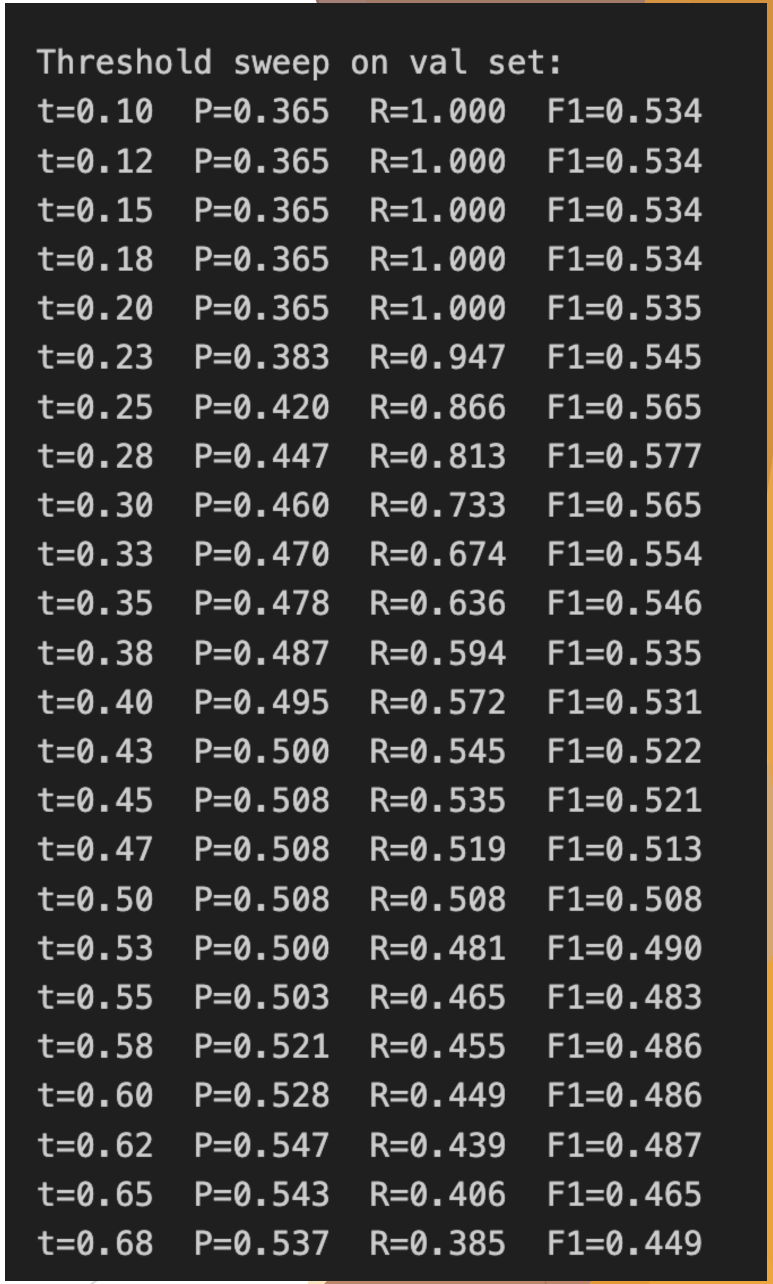
Threshold tuning : 0.25 - F1 = 56.5% , P = 42%, R = 87% (plagarism)

Threshold tuning : 0.45

Gird Search over Learning Rate and Epochs

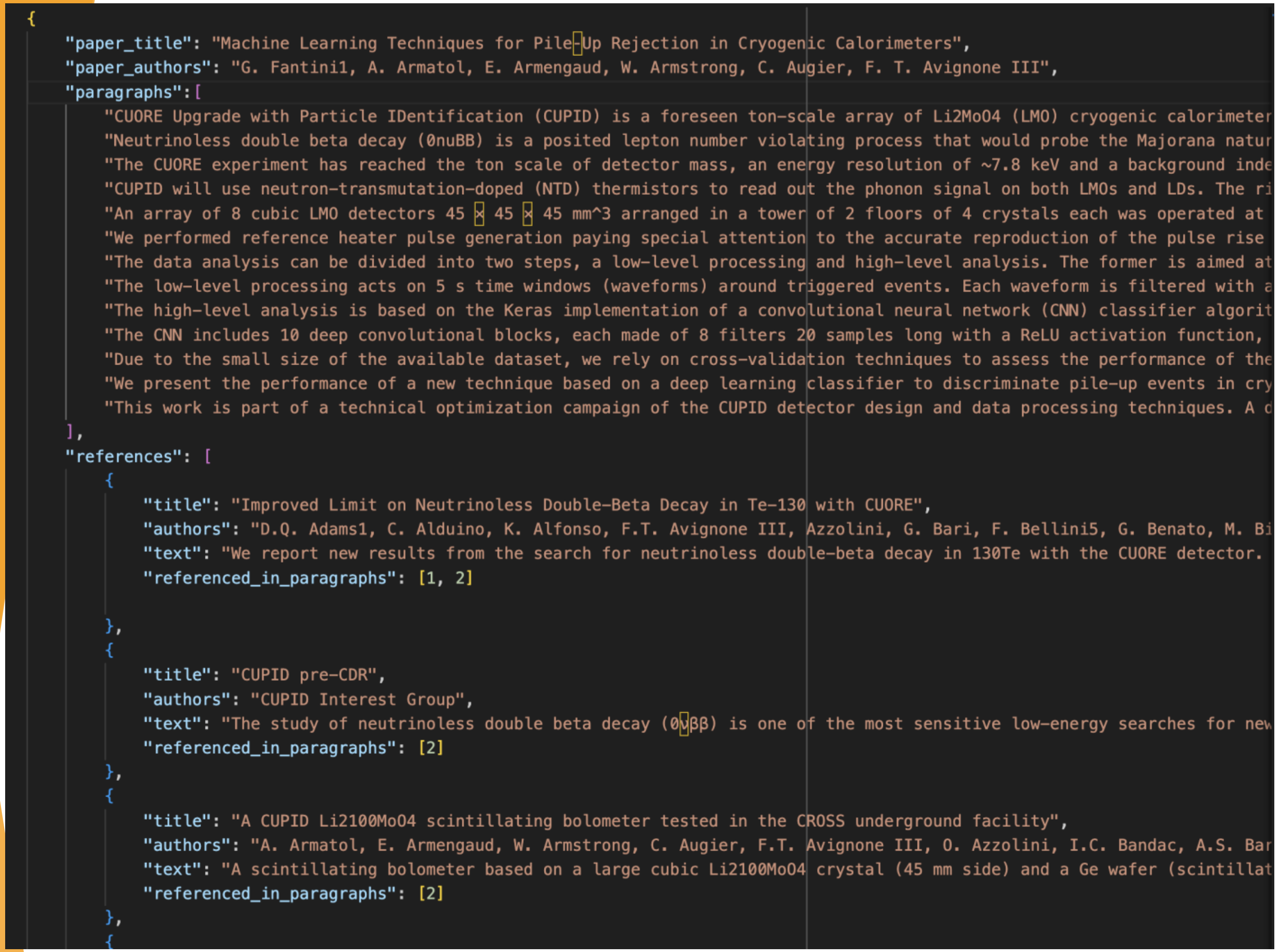
Final Model Metrics:

Epoch = 4
Learning Rate = 1e-05
Threshold = 0.45
Val = acc: 62%, prec: 48%, recall = 56%, f1 = 52%
Test = acc: 64%, prec: 46%, recall = 63%, f1 = 53%



```
Threshold sweep on val set:
t=0.10  P=0.365  R=1.000  F1=0.534
t=0.12  P=0.365  R=1.000  F1=0.534
t=0.15  P=0.365  R=1.000  F1=0.534
t=0.18  P=0.365  R=1.000  F1=0.534
t=0.20  P=0.365  R=1.000  F1=0.535
t=0.23  P=0.383  R=0.947  F1=0.545
t=0.25  P=0.420  R=0.866  F1=0.565
t=0.28  P=0.447  R=0.813  F1=0.577
t=0.30  P=0.460  R=0.733  F1=0.565
t=0.33  P=0.470  R=0.674  F1=0.554
t=0.35  P=0.478  R=0.636  F1=0.546
t=0.38  P=0.487  R=0.594  F1=0.535
t=0.40  P=0.495  R=0.572  F1=0.531
t=0.43  P=0.500  R=0.545  F1=0.522
t=0.45  P=0.508  R=0.535  F1=0.521
t=0.47  P=0.508  R=0.519  F1=0.513
t=0.50  P=0.508  R=0.508  F1=0.508
t=0.53  P=0.500  R=0.481  F1=0.490
t=0.55  P=0.503  R=0.465  F1=0.483
t=0.58  P=0.521  R=0.455  F1=0.486
t=0.60  P=0.528  R=0.449  F1=0.486
t=0.62  P=0.547  R=0.439  F1=0.487
t=0.65  P=0.543  R=0.406  F1=0.465
t=0.68  P=0.537  R=0.385  F1=0.449
```

Inference data

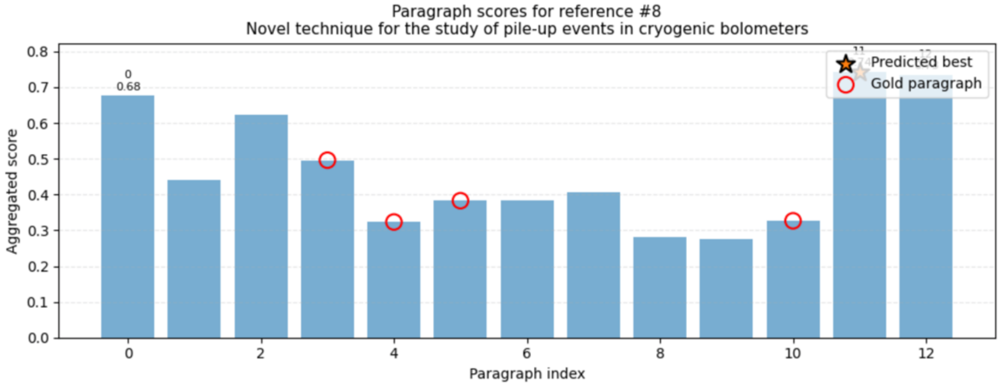


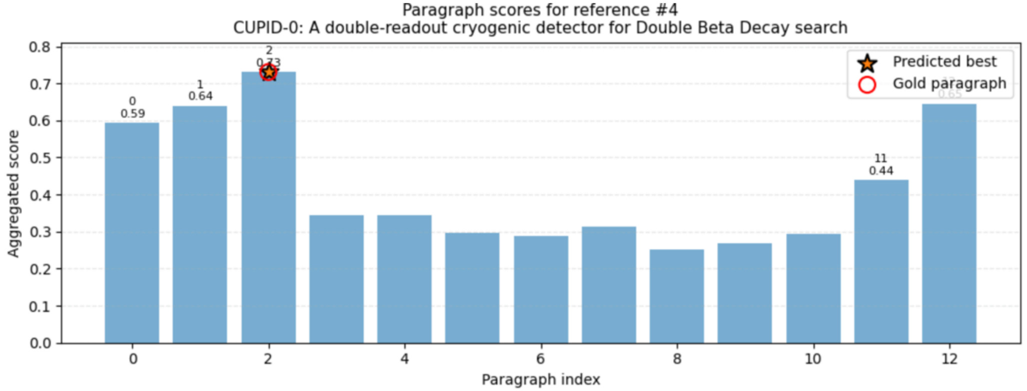
```
{
    "paper_title": "Machine Learning Techniques for Pile‐Up Rejection in Cryogenic Calorimeters",
    "paper_authors": "G. Fantini1, A. Armatol, E. Armengaud, W. Armstrong, C. Augier, F. T. Avignone III",
    "paragraphs":[
        "CUORE Upgrade with Particle IDentification (CUPID) is a foreseen ton-scale array of Li2MoO4 (LMO) cryogenic calorimeter
        "Neutrinoless double beta decay (0nuBB) is a posited lepton number violating process that would probe the Majorana natur
        "The CUORE experiment has reached the ton scale of detector mass, an energy resolution of ~7.8 keV and a background inde
        "CUPID will use neutron-transmutation-doped (NTD) thermistors to read out the phonon signal on both LMOs and LDs. The ri
        "An array of 8 cubic LMO detectors 45 × 45 × 45 mm^3 arranged in a tower of 2 floors of 4 crystals each was operated at
        "We performed reference heater pulse generation paying special attention to the accurate reproduction of the pulse rise
        "The data analysis can be divided into two steps, a low-level processing and high-level analysis. The former is aimed at
        "The low-level processing acts on 5 s time windows (waveforms) around triggered events. Each waveform is filtered with a
        "The high-level analysis is based on the Keras implementation of a convolutional neural network (CNN) classifier algorit
        "The CNN includes 10 deep convolutional blocks, each made of 8 filters 20 samples long with a ReLU activation function,
        "Due to the small size of the available dataset, we rely on cross-validation techniques to assess the performance of the
        "We present the performance of a new technique based on a deep learning classifier to discriminate pile-up events in cry
        "This work is part of a technical optimization campaign of the CUPID detector design and data processing techniques. A d
    ],
    "references": [
        {
            "title": "Improved Limit on Neutrinoless Double-Beta Decay in Te-130 with CUORE",
            "authors": "D.Q. Adams1, C. Alduino, K. Alfonso, F.T. Avignone III, Azzolini, G. Bari, F. Bellini5, G. Benato, M. Bi
            "text": "We report new results from the search for neutrinoless double-beta decay in 130Te with the CUORE detector.
            "referenced_in_paragraphs": [1, 2]

        },
        {
            "title": "CUPID pre-CDR",
            "authors": "CUPID Interest Group",
            "text": "The study of neutrinoless double beta decay (0νββ) is one of the most sensitive low-energy searches for new
            "referenced_in_paragraphs": [2]
        },
        {
            "title": "A CUPID Li2100MoO4 scintillating bolometer tested in the CROSS underground facility",
            "authors": "A. Armatol, E. Armengaud, W. Armstrong, C. Augier, F.T. Avignone III, O. Azzolini, I.C. Bandac, A.S. Bar
            "text": "A scintillating bolometer based on a large cubic Li2100MoO4 crystal (45 mm side) and a Ge wafer (scintillat
            "referenced_in_paragraphs": [2]
        },
        {
```

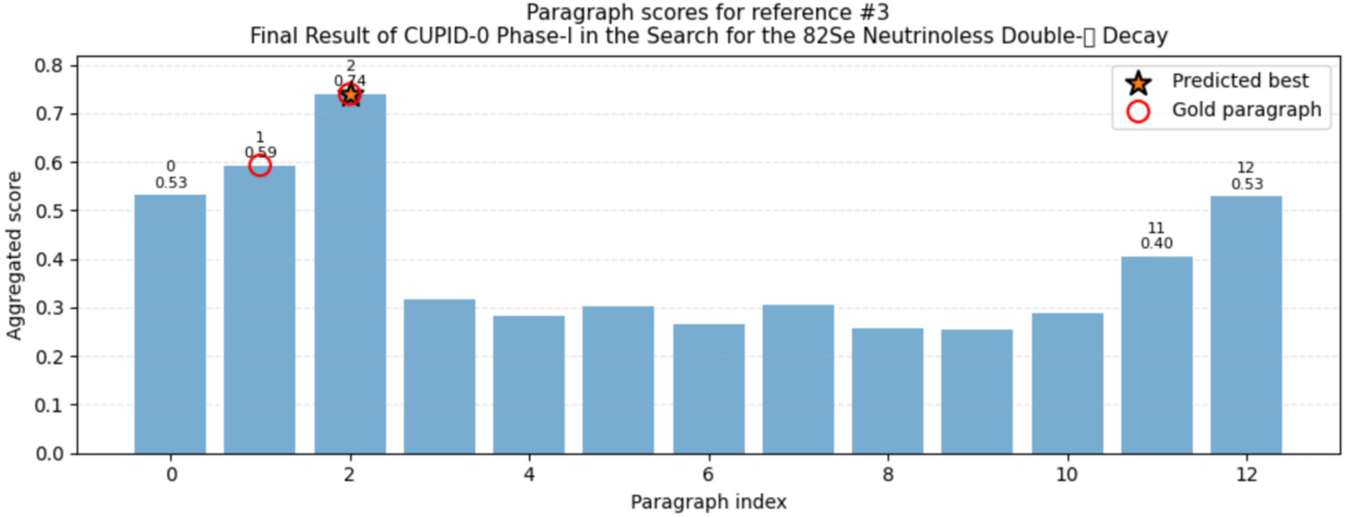
Inference!!!

1. Load fine-tuned SciBERT cross-encoder
2. Build Ref block
   - Title + author + text
3. Sentence lvl Scoring -> mean agg for paragraph lvl scoring
4. Rank top k paragraphs
5. Top 1 paragraph acc: 61.5% 8/13
6. Top 2 Paragraph acc: 77%

Result Examples


Paragraph scores for reference #8
Novel technique for the study of pile-up events in cryogenic bolometers


Paragraph scores for reference #4
CUPID-0: A double-readout cryogenic detector for Double Beta Decay search


Paragraph scores for reference #3
Final Result of CUPID-0 Phase-I in the Search for the 82Se Neutrinoless Double-▯ Decay

Limitations

- The S2ORC Dataset is huge (500 Gb +)
  - The dataset had to be stored, as there is no available API like there is for S2AG (Semantic Scholar Academic Graph, does not contain full text)
  - Used CJ’s personal computer to store the dataset
  - Working with the dataset took a long time (finding a valid paper for our dataset took on average 5-6 minutes, building 4000 rows took over 2 weeks)
- Corruption…
  - CJ’s drive failed when building the dataset of test papers, cutting off access to our main data source
  - The drive seems irrecoverable (including all of the contained personal data) ☹
  - Example test paper was hand parsed, so we only have 1

Future Improvements

- Get More Data!
- Increase the amount of text taken from references
- Improve decision making process based on model probabilities