# A Review & Summary of ConExion: Concept Extraction with Large Language Models

Alireza Alizadeh

## Main Idea and its Importance

With the advancements in LLMs during the recent years and their ability to generalize to an increasing number tasks every day, one task which helps us tremendously in classification and indexing of written media is concept extraction. Although the classical methods are viable and have acceptable performance, they sometimes fail to focus on concepts adjacent to human intuition all the while concepts are statistically relevant.

LLMs aim to solve the issues mentioned above however with the growing number of models every day and a lack of standard benchmarks for concept extraction. LLMs' performance on concept extraction remained somewhat of a mystery.

This paper is of particular importance since it tries to normalize, standardize and automate the evaluation process for the concept extraction task using LLMs. It provides detailed statistics about the datasets and each model's performance. Furthermore, considering the expenses and the hardware needed to run and evaluate LLMs, the findings of this paper and the performance summary it provides on a long list of LLMs from different companies and with variable sizes is of profound value.

## Inputs & Outputs

The system—due to its reliance on LLMs—works with a black box nature with a fairly simple input/output system. Inputs are essentially prompts and outputs are the text/response generated by the LLM.

Prompts(inputs) can be zero-shot or have various levels of instruction tuning. Included prompt types are the following:

- Zero-Shot (ZS)
- Zero-Shot with more domain information
- Zero-Shot with situational context
- Zero-Shot with task description
- Few-Shot (FS)

Prompt templates and detailed variations used can be found on *Table 1* of the paper.

Outputs are simple comma separated lists of extracted or generated concepts based on the specified tasks.

## Datasets

The paper utilizes two standard datasets both provided by MIDAS Research Laboratory.

Semeval2017 a dataset for benchmarking keyphrase extraction and generation techniques from abstracts of English scientific articles. For more details about the dataset please refer the original paper - https://arxiv.org/abs/1704.02853.

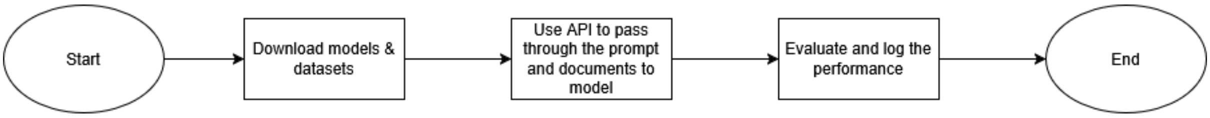
The Semeval-2017 dataset was originally proposed by *Isabelle Augenstein et al.* in the paper titled - SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications in the year 2017. The dataset consists of abstracts of 500 English scientific papers from the ScienceDirect open access publications. The selected articles were evenly distributed among the domains of Computer Science, Material Sciences and Physics. Each paper has a set of keyphrases annotated by student volunteers. Each paper was double-annotated, where the second annotation was done by an expert annotator. In case of disagreement, the annotations done by expert annotators were chosen. The original dataset was divided into train, dev and test splits, evenly distributed across the three domains. The train, dev and test splits had 350, 50 and 100 articles respectively.

Inspec a dataset for benchmarking keyphrase extraction and generation techniques from abstracts of English scientific papers. For more details about the dataset please refer the original paper - https://dl.acm.org/doi/pdf/10.3115/1119355.1119383.

The Inspec dataset was originally proposed by *Hulth* in the paper titled - Improved automatic keyword extraction given more linguistic knowledge in the year 2003. The dataset consists of abstracts of 2,000 English scientific papers from the Inspec database. The abstracts are from papers belonging to the scientific domains of *Computers and Control* and *Information Technology* published between 1998 to 2002. Each abstract has two sets of keyphrases annotated by professional indexers - *controlled* and *uncontrolled*. The *controlled* keyphrases are obtained from the Inspec thesaurus and therefore are often not present in the abstract's text. Only 18.1% of the *controlled* keyphrases are actually present in the abstract's text. The *uncontrolled* keyphrases are those selected by the indexers after reading the full-length scientific articles and 76.2% of them are present in the abstract's text. There is no information in the original paper about how these 2,000 scientific papers were selected. It is unknown whether the papers were randomly selected out of all the papers published between 1998-2002 in the *Computers and Control* and *Information Technology* domains or were there only 2,000 papers in this domain that were indexed by Inspec. The train, dev and test splits of the data were arbitrarily chosen.

## Implementation

The Implementation while being computationally expenses follows a fairly simple and straightforward logic. Models and datasets are downloaded to the server using shell scripts and then API calls are made to pass the prompt and document to the model and retrieve the response.


Start
Download models & datasets
Use API to pass through the prompt and documents to model
Evaluate and log the performance
End

## Results & Limitations

The results show that LLMs provide competitive performance compared to models specifically designed for concept extraction and in a best-case scenario would outperform them.

One limitation of this work is its reliance on exact lexical matching to filter concepts from the generated output. While this approach ensures that only terms present in the input document are retained, it fails to account for situations where an LLM generates semantically accurate concepts that do not have an exact match in the text. As a result, relevant concepts that capture the meaning are going to be discarded, negatively affecting recall and the overall evaluation.

A continuation of this paper can focus on multiple paths. For starters keeping the list of models and benchmarks updated with newly released models can be valuable. Furthermore, benchmarks and evaluation with domain specific datasets and more exploration into fine-tuning the models can be possible paths forward.