

# GPT-3 Abstract Classification for Scientometrics

This Jupyter book demonstrates how GPT-3 can be exploited to assign publications based on their abstracts to a commonly used scientific classification scheme, the Australian and New Zealand Standard Research Classification system (ANZSRC).

## Note

GPT-3 has been used to assist in creating and refining some of the code snippets and text paragraphs in this notebook.

The deep language model GPT-3, released to the public in 2022, has shown astounding performance in tasks that, from a human perspective, require a deeper understanding of the input texts. For historians, this potentially opens up the possibility of exploiting GPT-3's capability to approach larger corpora of source texts through distant reading. To showcase this potential and to provide a metric to estimate how well GPT-3 performs in one such a task, we test GPT-3's performance in assigning scientific publications to a particular scientific sub-field based on their abstracts.

As the dataset, we use metadata for publication items, including their abstracts, retrieved from the [Dimensions database](#) of scientific publications. Dimensions uses a standard scheme, the Australian and New Zealand Standard Research Classification system ([ANZSRC](#)), to classify all individual articles in its dataset. Dimensions' classification has been generated with the help of supervised machine learning based on extensive training sets and curated keyword searches. The approach taken by Dimensions in this classification is outlined on their [website](#), in their [Guide to the Dimensions Data Approach](#), and described in more detail in [Porter 2023](#).

In the following, we take Dimensions' field of research classification as the ground truth and test whether it can be reproduced with limited effort by exploiting the global knowledge of GPT-3 together with fine-tuning the language model to the task at hand. It is shown that GPT-3 can indeed be tuned to reproduce Dimensions' field of research classification extremely well. The results presented here lay the methodical foundation to apply GPT-3 to corpora, in particular historical ones, where, in contrast to the items in the Dimensions DB, an assignment to fields of research is lacking, opening such corpora up to scientometric analysis such as, for instance, in [Shtovba 2019](#). The success in reproducing the field of research assignment furthermore suggests that GPT-3 can likewise be used to extract other concepts of interest that have abstraction levels similar to those of scientific subfields (e.g. determining certain methods that have been used in a publication, etc.). On the downside GPT-3 is neither a free technology nor can fine-tuned models currently freely be shared between users. Not least because of this, one should think twice before using GPT-3 for scientific research purpose.

[Data](#)

[Light data analysis](#)

[Classification with GPT-3](#)

[Conclusion and outlook](#)

[Funding Information](#)