An Instruction Dataset for Extracting Quantum Cascade Laser Properties from Scientific Text Authors

Deperias Kerre, Anne Laurent, Kenneth Maussang, Dickson Odhiambo Owuor

Article Information

Article title

An Instruction Dataset for Extracting Quantum Cascade Laser Properties from Scientific Text

Authors

Deperias Kerre[a,c] *, Anne Laurent [a], Kenneth Maussang[b], Dickson Owuor [c]

Affiliations

[a] LIRMM, Univ Montpellier, CNRS, Montpellier, France

[b] Institut d'Electronique et des Systèmes, UMR 5214, Univ Montpellier, CNRS, Montpellier, France

[c] SCES, Strathmore University, Nairobi, Kenya

Corresponding author's email address and Twitter handle

*Corresponding author: Deperias Kerre, Email: dkerre@strathmore.edu

Keywords

Information Extraction, Large Language Models, Machine Learning, Quantum Cascade Lasers

Abstract

Quantum Cascade Lasers (QCL) are promising semiconductor lasers, compact and powerful, but of complex design. Availability of structured data of the QCL properties can support data mining activities that seek to understand the relationship between these properties, for instance between the design and performance features. The main open source of QCL data is in scientific text which in most cases is usually unstructured. One of the ways to extract and organize this data is by utilizing Information Extraction techniques. These techniques can accelerate the process of curating QCL properties data from scientific articles for further analysis. One of the main challenges in developing machine learning algorithms for extraction of QCL properties from text is lack of quality training data for these algorithms. Large Language Models (LLMs) have demonstrated great capabilities in materials property extraction from text. They however experience challenges with domain specific properties, for instance the heterostructure and design types in the QCL domain hence for adaptation. In this paper, we present an original instruction dataset for training and evaluation of large language models (LLMs) for QCL properties extraction from text. The data is generated by augmenting sample sentences from scientific articles with GPT-3.5 instruct with a few shot strategy. The dataset then is manually annotated with the help of QCL experts and is composed of 1300 rows of training examples consisting of an Instruction, Input Text and the Output.

SPECIFICATIONS TABLE

Subject	Computer Science, Materials Science
Specific subject area	Artificial Intelligence (Information Extraction and Text Mining)- Extraction of Quantum Cascade Laser properties from text.

Type of data	Processed data in CSV format.
Data collection	The dataset was generated by augmenting sample sentences from gold open access articles in the Quantum Cascade Laser domain. The sentences are augmented using the GPT-3.5 instruct model in order to generate similar sentences with various transformations such as rephrasing, new values of properties and generation of new sentences capturing same concepts but with different examples. The data is then labelled with an instruction to the large language model and the expected output for the property to be extracted.
Data source location	Sentences from 12 Gold open access articles from the AIP (American Institute of Physics) and IOP (Institute of Physics) publishers in accordance with the text and data mining policies. The articles document proposed Quantum Cascade Laser devices and their properties.
Data accessibility	Repository name: Recherche Data Gouv Data identification number: 10.57745/U3U7XR Direct URL to data: https://doi.org/10.57745/U3U7XR The dataset is licensed under the CC-BY license to enable re-use.

VALUE OF THE DATA

- The dataset provides a background in understanding the relationship between the various Quantum Cascade Laser properties. This is by enabling training of machine learning models to extract these properties from text to generate data that can be used for analysis for instance, in the prediction of the laser performance properties based on the design features.
- The dataset provides a benchmark dataset for evaluating large language models' performance for the task of information extraction of properties in the Quantum Cascade Laser domain. It covers the following properties: working temperature, frequency, power, the heterostructure stacking materials and design types. LLMs experience challenges in identifying QCL domain specific properties with direct prompting.
- The dataset gives a wide range of diverse sentences that capture the Quantum Cascade Laser properties in various ways hence enabling models to learn the various ways of expressing the properties. Most of the QCL properties are summarized in the abstract with the properties mentioned at sentence level. The dataset is therefore suitable for the extraction of properties from the abstracts. The focus is on articles describing a single novel QCL device hence enabling the generation of structured QCL properties data for various devices discussed in different papers.
- The dataset serves as central resource for researchers in the field of Information Extraction with an interest of developing efficient methodologies for materials science data extraction from scientific text. The dataset can be extended by further augmentation or with additional Quantum Cascade Laser properties to provide a diverse dataset for properties data extraction from text. The format of the dataset can also be adopted in developing information extraction datasets for other properties in materials science.

BACKGROUND

The main motivation behind the generation of this dataset is for understanding the relationship between the various properties of Quantum Cascade Lasers (QCLs). A Quantum Cascade Laser (QCL) is a semiconductor laser device made up of different layers of semiconductor materials and with a spectral emission range in the mid-to far infrared [1]. A QCL device with certain design characteristics (the materials layers and the design types) has corresponding working properties such as the temperature, power and frequency among others. This implies that a particular design has an influence on the working properties of the proposed QCL device. Understanding these relationships between properties can give insights on the fabrication of QCL devices with target properties. One of the ways to explore and understand the QCL data is by capturing it in knowledge representation systems such as Ontologies and Knowledge Graphs. There already exists ontologies for concepts such as units of materials properties such as the QUDT Ontology [2] and Ontologies for other characteristics of materials data such as measurable properties for instance the Materials Design Ontology [3]. There is also a proposal for the Ontology for the QCL properties in order to formally represent these properties [4]. With already such existing resources for representing the QCL properties, access to structured QCL properties data will accelerate the process of developing platforms for analysing the relationships between the QCL properties.

In most cases, the QCL design and working properties are captured in scientific literature. An example of a description of QCL laser properties can be given as “A Terahertz quantum cascade laser with a rather high injection coupling strength based on an indirectly pumped scheme is designed and experimentally implemented. To effectively suppress leakage current, the chosen quantum cascade module of the device is based on a five-well GaAs/Al0.25Ga0.75As structure. The device lases up to 151 K with a lasing frequency of 2.67 THz.”[5]. This text description contains various QCL properties for instance the heterostructure, working temperature, frequency and design type. Rule-based methods have been proposed for QCL properties extraction from text [6]. These are however limited to the reporting style in text and requires a lot of expertise and effort in rewriting of the rules when there is even a slight change in the text structure or being applied to other QCL properties.

The existence of QCL properties in scientific text necessitates the need for efficient methods to extract these properties to collect data for analysis. The CQL properties data constitutes domain specific data that cannot be efficiently mined by information extraction methods without adaptation. This calls for quality datasets that can be used in training these algorithms to efficiently extract these properties from scientific text. Based on this motivation, we propose an instruction dataset that provides an initial attempt in generating quality data for developing efficient algorithms for QCL property extraction from text [7].

DATA DESCRIPTION

The dataset consists of a CSV file with 1300 rows of sentences generated from sample sentences from scientific articles. The data is availed a single file without any partitions for instance the training, test or validation sets. The CSV file has three columns: Instruction, Input Text and Output. The Instruction column consists of sentences that entails an instruction to the large language model to extract a specific QCL property from the text. The input text column consists of sentences containing the QCL properties mentioned in them. The output column contains the expected output of extracted properties based on the user inquiry in the instruction column. From the machine learning perspective, the output column may be seen as the label for the input text and the instruction.

MATERIALS AND METHODS

Sentence Sampling

In order to obtain sample sentences for augmentation, we focus on scientific open access articles that solely describes originally proposed QCL devices and the corresponding properties . We don't consider articles that proposes applications of these devices in other applications. We also limit this on articles whose publishers permit text and data mining for research purposes. Through this process, we were able to extract 36 sentences from 12 scientific articles to be used in the augmentation process [5], [8-18]. The sentences contain the various QCL properties of interest: working temperature, power, laser frequency and the laser design types.

Dataset Generation

Large Language Models have shown great progress in text data augmentation [19,20]. Inspired by this, we employ a few shot strategy to generate an augmented dataset of QCL properties from sample sentences extracted from the scientific articles using GPT 3.5 instruct model. The sentences are parsed to the GPT-3.5 instruct model in a few short manner to paraphrase and generate similar sentences with varying expression styles and values for the QCL properties. The process is repeated in iterations hence obtaining different sentences with various values for the laser properties. The generated sentences are validated by an expert to make sure those that make sense are the only one included in the final dataset. The generated sentences are combined with the original sentences to make a total of 1300 rows. Figure 1. Dataset Generation Process shows the overall process of the dataset generation. The LLM prompts and the original sentences are publicly availed at the Gitlab repository: https://gite.lirmm.fr/dkerre/qpInstruct .

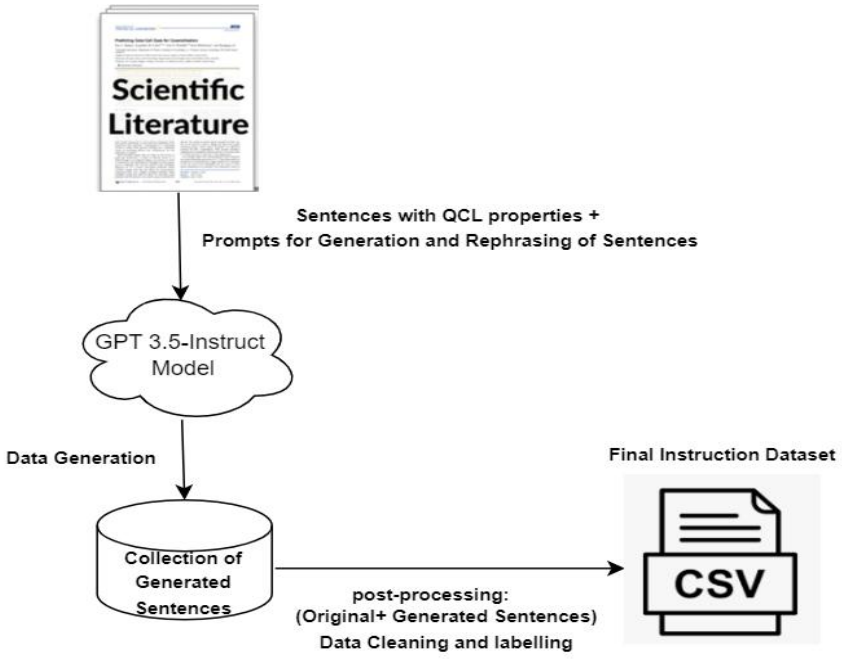

Figure 1. Dataset Generation Process

Data Annotation

The final stage entails annotating the data with various kinds of information to be extracted from the text. Two columns are added i.e. the Instruction column and the output column. In the instruction column, the instructions to the LLM models are captured. The output column consists of the expected

properties to be extracted as per the instruction in the Instruction column. The output column is labelled with assistance from the QCL domain experts.

LIMITATIONS

The dataset only captures the following QCL properties: the working temperature, power, frequency , design type and the heterostructure materials. It was also a challenge to generate more sentences due to the cost implications of using GPT and the required labelling and post-processing efforts. Researchers can however use this dataset to augment and generate larger instruction datasets for QCL properties extraction from text. The quality of the generated sentences may not fully represent the nature of descriptions given in scientific articles detailing the QCL properties.

ETHICS STATEMENT

The authors have read and followed the ethical requirements for publication in Data in Brief. The current work does not involve human subjects, animal experiments, or any data collected from social media platforms. The scientific articles were sourced with permission from the publishers and with adherence to the text and data mining policies of the publishers.

CREDIT AUTHOR STATEMENT

Deperias Kerre: Resources, Investigation, Methodology, Data Preparation, Writing-Original Draft, Writing-Review and Editing

Anne Laurent: Methodology, Writing- Review and Editing, Supervision

Kenneth Maussang: Methodology, Writing- Review and Editing, Supervision

Dickson Owuor: Methodology, Writing- Review and Editing, Supervision

ACKNOWLEDGEMENTS

This study has been funded by the French Embassy in Kenya (Scientific and Academic Cooperation Department) and the CNRS (under the framework “Dispositif de Soutien aux Collaborations avec l’Afrique sub-saharienne”). The authors would also like to appreciate Strathmore university, School of Computing and Engineering sciences and the Doctoral academy for creating an opportunity for this work to be realized.

DECLARATION OF COMPETING INTERESTS

- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Faist, J., Capasso, F., Sivco, D. L., Hutchinson, A. L., Sirtori, C., & Cho, A. Y. (1995). Quantum cascade laser: a new optical source in the mid-infrared. Infrared Physics & Technology, 36(1), 99-103.
2. Haas, R., Keller, P., Hodges, J., and Spivak, J. (2014). Quantities, units, dimensions and data types ontologies (qudt). Technical report, Top Quadrant and NASA.
3. Li, H., Armiento, R., and Lambrix, P. (2020). An ontology for the materials design domain. In The Semantic Web – ISWC 2020, page 212–227.
4. Kerre, D., Laurent, A., Maussang, K., and Owuor, D. (2023). A concise ontological model of the design and optoelectronic properties in the quantum cascade laser domain. preprint.

5. Razavipour, S. G., Dupont, E., Chan, C. W. I., Xu, C., Wasilewski, Z. R., Laframboise, S. R., ... & Ban, D. (2014). A high carrier injection terahertz quantum cascade laser based on indirectly pumped scheme. *Applied Physics Letters*, *104*(4).
6. Kerre, D., Laurent, A., Maussang, K., & Owuor, D., (2023, August). A text mining pipeline for mining the quantum cascade laser properties. In European Conference on Advances in Databases and Information Systems (pp. 393-406). Cham: Springer Nature Switzerland.
7. Kerre, D., Laurent, A., Maussang, K., & Owuor, D., (2024). An Instruction Dataset for Extracting Quantum Cascade Laser Properties from Scientific Text. Recherche Data Gouv: DOI: 10.57745/U3U7XR.
8. Kumar, S., Hu, Q., & Reno, J. L. (2009). 186 K operation of terahertz quantum-cascade lasers based on a diagonal design. *Applied Physics Letters*, *94*(13).
9. Williams, B. S., Callebaut, H., Kumar, S., Hu, Q., & Reno, J. L. (2003). 3.4-THz quantum cascade laser based on longitudinal-optical-phonon scattering for depopulation. *Applied Physics Letters*, *82*(7), 1015-1017.
10. Hempel, M., Röben, B., Niehle, M., Schrottke, L., Trampert, A., & Grahn, H. T. (2017). Continuous tuning of two-section, single-mode terahertz quantum-cascade lasers by fiber-coupled, near-infrared illumination. *AIP Advances*, *7*(5).
11. Wang, F., Slivken, S., Wu, D. H., Lu, Q. Y., & Razeghi, M. (2020). Continuous wave quantum cascade lasers with 5.6 W output power at room temperature and 41% wall-plug efficiency in cryogenic operation. *AIP Advances*, *10*(5).
12. Wang, T., Liu, J. Q., Chen, J. Y., Liu, Y. H., Liu, F. Q., Wang, L. J., & Wang, Z. G. (2013). Continuous-wave operation of terahertz quantum cascade lasers at 3.2 THz. *Chinese Physics Letters*, *30*(6), 064201.
13. Lü, X., Röben, B., Schrottke, L., Biermann, K., & Grahn, H. T. (2021). Correlation between frequency and location on the wafer for terahertz quantum-cascade lasers. *Semiconductor Science and Technology*, *36*(3), 035012.
14. Khabibullin, R. A., Shchavruk, N. V., Pavlov, A. Y., Klochkov, A. N., Glinskiy, I. A., Tomosh, K. N., ... & Zhukov, A. E. (2019). Design and fabrication of terahertz quantum cascade laser with double metal waveguide based on multilayer GaAs/AlGaAs heterostructures. In *IOP Conference Series: Materials Science and Engineering* (Vol. 475, No. 1, p. 012020). IOP Publishing.
15. Ohtani, K., Turčinková, D., Bonzon, C., Benea-Chelmus, I. C., Beck, M., Faist, J., ... & Stutzki, J. (2016). High performance 4.7 THz GaAs quantum cascade lasers based on four quantum wells. *New Journal of Physics*, *18*(12), 123004.
16. Wang, X., Shen, C., Jiang, T., Zhan, Z., Deng, Q., Li, W., ... & Duan, S. (2016). High-power terahertz quantum cascade lasers with~ 0.23 W in continuous wave mode. *Aip Advances*, *6*(7).
17. Valmorra, F., Scalari, G., Ohtani, K., Beck, M., & Faist, J. (2015). InGaAs/AlInGaAs THz quantum cascade lasers operating up to 195 K in strong magnetic field. *New Journal of Physics*, *17*(2), 023050.
18. Kumar, S., Williams, B. S., Hu, Q., & Reno, J. L. (2006). 1.9 THz quantum-cascade lasers with one-well injector. *Applied Physics Letters*, *88*(12).
19. Fang, L., Lee, G. G., & Zhai, X. (2023). Using gpt-4 to augment unbalanced data for automatic scoring. arXiv preprint arXiv:2310.18365.
20. Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., ... & Li, X. (2023). Auggpt: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007.