# NL4Opt Competition: Formulating Optimization Problems Based on Their Natural Language Descriptions

**Rindranirina Ramamonjison**[1] RINDRANIRINA.RAMAMONJISON@HUAWEI.COM
**Timothy T. Yu**[1] TIMOTHYT.YU@HUAWEI.COM
**Raymond Li**[2] RAYMONDL@CS.UBC.CA
**Haley Li**[1,2] HALEYLI@CS.UBC.CA
**Giuseppe Carenini**[2] CARENINI@CS.UBC.CA
**Bissan Ghaddar**[3] BGHADDAR@IVEY.CA
**Shiqi He**[1,2] SHIQIHE@CS.UBC.CA
**Mahdi Mostajabdaveh**[1] MAHDI.MOSTAJABDAVEH1@HUAWEI.COM
**Amin Banitalebi-Dehkordi**[1] AMIN.BANITALEBI@GMAIL.COM
**Zirui Zhou**[1] ZIRUI.ZHOU@HUAWEI.COM
**Yong Zhang**[1] YONG.ZHANG3@HUAWEI.COM

[1] *Huawei Technologies Canada, Burnaby, Canada*
[2] *University of British Columbia, Vancouver, Canada*
[3] *Ivey Business School, London, Canada*

**Editors:** Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

## Abstract

The Natural Language for Optimization (NL4Opt) Competition was created to investigate methods of extracting the meaning and formulation of an optimization problem based on its text description. Specifically, the goal of the competition is to increase the accessibility and usability of optimization solvers by allowing non-experts to interface with them using natural language. We separate this challenging goal into two sub-tasks: (1) recognize and label the semantic entities that correspond to the components of the optimization problem; (2) generate a meaning representation (i.e. a logical form) of the problem from its detected problem entities. The first task aims to reduce ambiguity by detecting and tagging the entities of the optimization problems. The second task creates an intermediate representation of the linear programming (LP) problem that is converted into a format that can be used by commercial solvers. In this report, we present the LP word problem dataset and shared tasks for the NeurIPS 2022 competition. Furthermore, we present the winning solutions. Through this competition, we hope to bring interest towards the development of novel machine learning applications and datasets for optimization modeling.

**Keywords:** Operations Research, NLP, Entity Recognition, Semantic Parsing, Math Word Problems, Controllable Generation, ChatGPT Comparison, Linear Programming

## 1. Introduction

Operations research (OR) tools can be leveraged to model and solve many real-world decision-making problems analytically and efficiently. OR is a field of applied mathematics that has been proven beneficial in many applications such as supply chain management (Maloni and Benton, 1997), production planning (Pochet and Wolsey, 2006), bike-share ridership and efficiency in urban cities (Beairsto et al., 2021; Ma et al., 2016), managing wastewater collection and treatment systems (Tao et al., 2020), and finding a revenue-maximizing pricing strategy for businesses (Bitran and Caldentey, 2016). Different types of optimization problems can be solved using standard optimization algorithms such as the simplex (Nash, 2000) or interior-point method (Karmarkar, 1984). However, modeling real-world problems into proper formulations as input to optimization solvers is still an iterative and strenuous process. First, the problem must be described by the stakeholder in the language of a domain expert. Then, an OR expert must extract the decision variables, objective, and the constraints from the description. Finally, the problem must be re-written in an algebraic modeling language that solvers can interpret.

Through the NL4Opt Competition, we investigate the feasibility of learning-based natural language interfaces for optimization solvers. To do so, we explored the practicality of partially automating the formulation of optimization problems. In particular, semantic parsing is a general task for extracting machine-interpretable meaning representations from natural language utterances. They have been well-studied for designing NLP systems that interact with database systems (Zhong et al., 2017; Gan et al., 2020), Unix machines (Lin et al., 2018), knowledge base systems (Berant and Liang, 2014; Dong and Lapata, 2016) or dialog systems (Guo et al., 2018). However, extracting the formulation of optimization problems is still an under-explored problem. Meanwhile, solving math word problems with NLP has seen sustained research activity (Koncel-Kedziorski et al., 2016; Hopkins et al., 2019; Miao et al., 2020; Patel et al., 2021) with researchers focused on finding the correct answers to elementary algebraic and arithmetic problems. In contrast, rather than exploring methods of producing the solution to the problem, we focus on converting optimization problems into a form that can be passed to commercial optimization solvers to efficiently find optimal solutions.

Lately, a few related challenges have been created for analyzing scientific texts. For instance, Harper et al. (2021) proposed the MeasEval challenge focused on extracting counts and measurements from clinical documents and finding the attributes of those quantities. Another popular challenge was MultiCoNER (Malmasi et al., 2022) which focused on detecting semantically ambiguous and complex entities from documents written in 11 languages spanning 13 tracks. The NL4Opt Competition expands this task by not only detecting complex entities from optimization problems but also generating the equivalent mathematical formulation.

## 2. The NL4Opt Competition

The NL4Opt Competition explores the design of natural language interfaces for optimization solvers. The results of this competition forward the accessibility and usability of these solvers and allow non-experts to solve important problems from various industries. Specifically, we used this competition to explore methods of converting a natural language

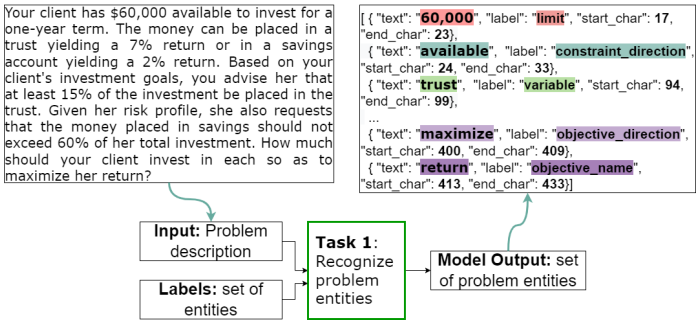


Figure 1: Description of dataset for sub-task 1: "recognizing the problem entities"

description of an optimization problem into a mathematical formulation. This goal was separated into two inter-related sub-tasks:

1. Recognition of optimization problem entities,

2. Generation of problem formulation.

The first sub-task was to recognize optimization model entity types (i.e., constraint direction, constraint limit, objective direction, objective name, parameter, variable) from the problem description. In the first sub-task, the goal was to detect text spans from the problem description that represent semantic entities of the optimization problem and to tag them according to the listed entity types. This sub-task aimed to reduce the ambiguity by identifying important components of the optimization model. An illustration of sub-task 1 is provided in Figure 1.

The second sub-task was to generate a precise meaning representation of the optimization formulation. This sub-task was simplified using the ground truth information of the problem entities from the first sub-task. An illustration of sub-task 2 is provided in Figure 2.

The proposed sub-tasks are characterized by the following challenges:

1. **Unstructured multi-sentence input**. An optimization description is the input document that describes the decision variables, objective, and a set of constraints. In addition, the structure of the input varies depending on the structure of the optimization problem and the linguistic style. Thus, the multi-sentence input exhibits a high level of compositionality and ambiguity due to the variability of the linguistic patterns, of the problem domains, and of the problem structures.

2. **Mismatched inputs and outputs**. The contextual information from the input description is abstracted away in the target formulation. Therefore, the absence of contextual clues in the output makes it difficult to align the input-output pair. Thus, the meaning representation of the problem formulation (i.e. the output of generation model) is important as it bridges the problem description and the mathematical formulation. In fact, semantically equivalent representations may have syntactically different forms and can lead to different performance (Guo et al., 2020).

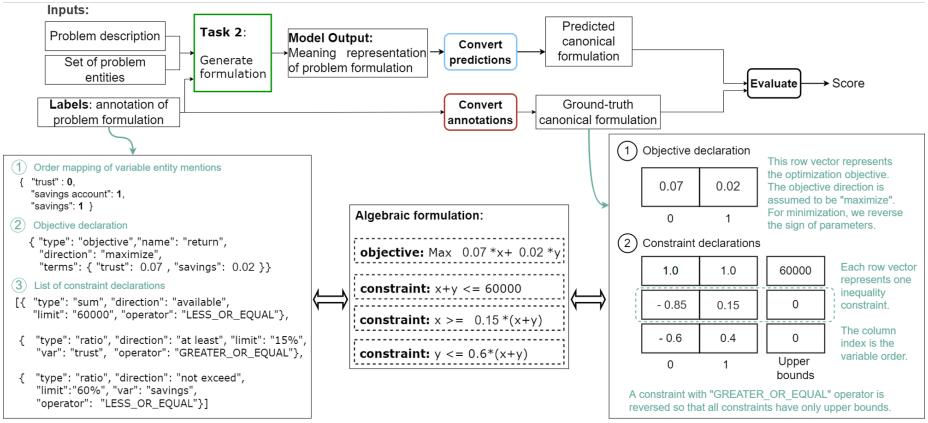


Figure 2: Dataset annotation and evaluation protocol for sub-task 2: "generating the problem formulation"

3. **Low-resource learning constraint**. Specialized knowledge is required to create a dataset thereby drastically increasing the cost of dataset creation. The design of machine learning models for this task is challenging as they must learn from a small number of expert-annotated examples.

4. **Domain-agnostic parsing**. Finally, OR tools are applied to disparate problem domains (e.g. forestry, transportation or medicine) (Williams, 2013). As a result, the learning-based solution must generalize not only to new problem instances but also to new application domains.

### 2.1. Evaluation

**NER sub-task:** we evaluated the models based on their achieved micro-averaged F1 score given by:

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R}, \tag{1}$$

where $P$ and $R$ are the average precision and average recall further averaged over all entity types, respectively.

**Generation sub-task:** we used an application-specific metric since the task was motivated by the need to precisely formulate the optimization problem. We have benchmarked the models based on the declaration-level mapping accuracy given by:

$$\mathrm{Acc} = 1 - \frac{\sum_{i=1}^{N} \min\left\{\mathrm{FP}_i + \mathrm{FN}_i, D_i\right\}}{\sum_{i=1}^{N} D_i}, \tag{2}$$

where $N$ is the number of linear programming word problems (LPWPs) in the test set. For each LPWP $i$, $D_i$ is the number of ground-truth declarations. The false positive $\mathrm{FP}_i$ is the number of non-matched predicted declarations whereas the false negative $\mathrm{FN}_i$ denotes the number of ground-truth declarations without a match. To clarify the evaluation protocol,

we have emphasized that the canonical representation, as described in Figure 2, would be used to compare the ground-truth and predicted formulations.

### 2.2. Competition statistics

Over 150 teams registered for this competition combining for a total of more than 300 valid independent submissions. The demographics of registered participants were affiliated as follows: 30% post-secondary, 30% independent, 30% unspecified, and 10% industry. There were 19 teams with valid submissions to sub-task 1 and 9 to sub-task 2. These teams reported the following affiliations: 60% industry, 25% post-secondary, and 15% independent.

### 2.3. Additional Competition Details

All the details and relevant information of the competition are made accessible at the competition website (https://nl4opt.github.io/). This website contains the rules, a FAQ/Tutorial section, access to the starter kit, and final results of the competition. We have released the test set and encourage all new and returning participants to leverage this competition as a benchmark for new methods. We especially welcome those that are interested in tackling the challenges listed above (i., unstructured input, misaligned input-output pair, low-resource learning constraint, generalizability).

## 3. The NL4Opt Dataset

**Dataset description.** A total of 1101 annotated LPWPs from 6 different domains were created for the NL4Opt Competition. We separated the dataset into 713, 99, and 289 samples for training, development, and testing, respectively. The data samples were distributed identically for both sub-tasks. It is important to evaluate submissions for generalizability towards unseen problem domains. Therefore, we include LPWPs from the three similar (source) domains of sales, advertising, and investment in both training, development and test splits. However, problems from three other (target) domains (production, transportation, and sciences) have been reserved for the development and test splits. To ensure that the development set was never used for training, we reviewed the final submitted code and re-trained all submissions prior to announcing the winning teams. Table 1 presents the number of samples and the ratio between the source and target domains for the three splits of data. An example of data and its annotations for the two sub-tasks is illustrated in Figures 1 and 2.

Table 1: Data Distribution.

| Split | #/samples | source:target |
|---|---|---|
| Train | 713 | 1:0 |
| Dev | 99 | 1:3 |
| Test | 289 | 1:3 |

For the first sub-task, the input is the problem description and the output is the set of entities that correspond to the components of the problem (Figure 1). The entities are

labelled according to predefined entity types. The labels were provided in Spacy format and in BIO tagging format.

For the second sub-task, the inputs are the problem description, its set of problem entities, and the order mapping of variable mentions. The ground-truth label annotations consist of the objective declaration and the constraints declarations as shown in Figure 2. The output of the semantic parser is the meaning representation of those declarations. As shown in Figure 2, the meaning representation should be converted to a canonical form for evaluation. Participants were encouraged to either design their own meaning representation or use the representation and conversion scripts from our pilot study.

**Dataset creation.** A team of 20 AI engineers and OR experts spent three months to create our preliminary LPWP dataset. This team used the Prodigy tool (Montani and Honnibal, 2018) to manually create and annotate this preliminary dataset containing 600 problems. Within the team, five were tasked with verifying that each problem adhered to specific guidelines to ensure diversity in problem types and language patterns. Throughout the process of creating the remaining 501 samples, suggested annotations were generated using a preliminary NER model trained on the preliminary dataset. For the second sub-task, we created a custom Prodigy recipe and a Python script to efficiently annotate the ground-truth declarations of the objective and constraints. All of the new problems and annotations for both sub-tasks were verified and corrected by at least two experts. Ramamonjison et al. (2022) describes in more details the data creation process (e.g., exclusion criteria, inter-annotator agreement, correction process, average duration of each step, etc.).

Note that we did not use existing datasets from third parties and have released the dataset[1] under the MIT License to benefit the research community.

## 4. Baseline Models

Participants had access to the code base from our pilot study that is described in more details in Ramamonjison et al. (2022). Most participants built upon this by implementing their own methods on the provided code base.

### 4.1. Sub-task 1

The starter kit for sub-task 1 can be found in the NL4Opt repository[2]. The baseline model, XLM-RoBERTa-base (XLM-R-base) (Conneau et al., 2019), was trained and fine-tuned by minimizing the log-likelihood loss. As part of the pilot study, we reported[3] the baseline model's performance on the test set when evaluated on the source domain, target domain, and entire test set for all entity types (i.e., constraint direction, limit, etc.). Based on this preliminary analysis, the objective name was the most difficult to identify potentially due to its ambiguity. We expect the greatest improvements would arise from methods that are capable of accurately recognizing the objective names and their spans. **Evaluation:** *This baseline achieved an F1 score of* ***0.906*** *on the test split.*

1. All data are available at: https://github.com/nl4opt/nl4opt-competition
2. Sub-task 1 baseline is available at: https://github.com/nl4opt/nl4opt-subtask1-baseline
3. Stratified performance: https://github.com/nl4opt/nl4opt-subtask1-baseline/tree/main/baseline#results

### 4.2. Sub-task 2

The starter kit for sub-task 2 can be found in the NL4Opt repository[4]. The starter kit for sub-task 2 contains code to parse the XML-like intermediate representations and annotated examples into our Problem Formulation dataclass and code to score the submission. Additional information regarding the canonical representation, parsing, and scoring can be found in this notebook. For the generation sub-task, the baseline model is a BART encoder-decoder (Lewis et al., 2019) that leverages a prompt-guided generation and a copy mechanism to generate a meaning representation of the optimization formulation. **Evaluation:** *This baseline achieved an accuracy of* ***0.610*** *on the test set.*

## 5. Solutions

Table 2: Sub-task 1 winning results.

| Rank | Team | F1 score |
|---|---|---|
| 1 | Infrrd AI Lab | 0.939 |
| 2 | mcmc | 0.933 |
| 3 | PingAn-zhiniao | 0.932 |
| 4 | Long | 0.931 |
| 5 | VTCC-NLP | 0.929 |
| - | Baseline | 0.906 |

Table 3: Sub-task 2 winning results.

| Rank | Team | F1 score |
|---|---|---|
| 1 | UIUC-NLP | 0.899 |
| 2 | Sjang | 0.878 |
| 3 | Long | 0.867 |
| 4 | PingAn-zhiniao | 0.866 |
| 5 | Infrrd AI Lab | 0.780 |
| - | Baseline | 0.610 |

### 5.1. Sub-task 1

#### 5.1.1. First Place: Team Infrrd AI Lab

Team Infrrd AI Lab (JiangLong He, Mamatha N., Shiv Vignesh, Deepak Kumar, Akshay Uppal) leveraged **ensemble learning** with **augmentation** to achieve an F1 score of **0.939** on the test set. Their base model consists of text embedding, BiLSTM, and CRF layers. Through ablations studies, they found that the *majority voting* of an ensemble of 5 different models that were designed in a combination of XLM-R-base and RoBERTa-base transformers for text embeddings, BiLSTM layers, and CRF layers performed the best on the test set. They also implemented 4 types of data augmentation techniques during training. Namely, label-wise token replacement, synonym replacement, mention replacement, and shuffle within segments. For more details, refer to (He et al., 2022).

#### 5.1.2. Second Place: Team mcmc

Team mcmc (Kangxu Wang, Ze Chen, Jiewen Zheng) trained models for **ensemble learning** with **adversarial attacks** to achieve an F1 score of **0.933** on the test set. They found that implementing adversarial attack using the FGM proposed by Goodfellow et al. (2014) on the DeBERTa-large transformer (He et al., 2021) with a CRF layer performed the best on the development set. They trained 9 variations of this model using different random

4. Sub-task 2 baseline is available at: https://github.com/nl4opt/nl4opt-subtask2-baseline

initializations to form their ensemble and leveraged *majority voting* for the final prediction. For more details, refer to (Wang et al., 2023).

#### 5.1.3. Third Place: Team PingAn-zhiniao

Team PingAn-zhiniao (Qi Zeng, Xiuyuan Yang, Yixiu Wang, Chang Shu) augmented the fine-tuning process of the XLM-R-large transformer by implementing a **global pointer decoder** followed by a **multi-head decoder** to achieve an F1 score of **0.932** on the test set. This was the only sub-task 1 winning submission that did not use ensemble learning. Initially, they fine-tuned using the XLM-R-large encoder to produce the embeddings which was fed into both the global pointer decoder and multi-head decoder. Upon reaching an F1 score of 0.9 on the development set, the global pointer decoder was removed while the encoder with multi-head decoder model continued training.

#### 5.1.4. Fourth Place: Team Long

Team Long (Yuting Ning, Jiayu Liu, Longhu Qin, Tong Xiao, Shangzi Xue, Zhenya Huang, Qi Liu, Enhong Chen, Jinze Wu) leveraged **ensemble learning**, **adversarial training**, and some **post-processing** techniques to achieve an F1 score of **0.931** on the test set. They used XLM-R as the base model and leverage projected gradient descent method (Madry et al., 2017) and FGM for adversarial training. Augmentations included variables swapping, synonym replacement in objective names, and randomizing of numbers. They also implemented some quick-check rules to enforce consistency in tagging entity spans. Four models (XLM-R-base and XLM-R-large) were optimized for specific entity types and the final prediction was obtained through an emsemble learning framework. For more details, refer to (Ning et al., 2023) or their code[5].

#### 5.1.5. Fifth Place: Team VTCC-NLP

Team VTCC-NLP (Xuan-Dung Doan) proposed **ensemble learning** to achieve an F1 score of **0.929** on the test set. They also explored the use of ELMo embedding (Peters et al., 2018) and GCN models (Yao et al., 2018) and found that both improved the performance of the baseline model accuracy, but negatively impacted the performance when included for ensemble learning. The final ensemble consisted of XLMR, DeBERTaV3, and BART. For more details, refer to (Doan, 2022).

### 5.2. Sub-task 2

#### 5.2.1. First Place: Team UIUC-NLP

Team UIUC-NLP (Neeraj Gangwar, Nickvash Kani) **tagged the input** and implemented a **"decode all-at-once" strategy** to achieve an accuracy of **0.899** on the reserved test set. They used the BART-large encoder-decoder model and enriched the input by surrounding entities with XML-like tagging. Through ablation studies, they found the best performance when combining this input tagging strategy with generating all objective and constraint declarations at once. This team also reports higher sensitivity to hyperparameters and

5. Team Long code: https://github.com/bigdata-ustc/nl4opt

initial seeds when using the large version of BART compared to the base version. For more details, refer to (Gangwar and Kani, 2022) or their code[6].

#### 5.2.2. Second Place: Team Sjang

Team Sjang (Sanghwan Jang) used a **scaling hyperparameter** to introduce **entity tag embeddings** and they implement simple **data augmentation** to achieve an accuracy of **0.878** on the reserved test set. Compared to the baseline, they report a 16% increase in accuracy by implementing the BART-large model, a further 10% improvement by scaling the tag embedding, and another 1.5% through simple augmentations to the constraints by reversing the constraint direction. For more details, refer to (Jang, 2022) or their code[7].

#### 5.2.3. Third Place: Team Long

Team Long **redesigned the prompt**, implemented **data augmentation**, and leveraged **adversarial training** to achieve an accuracy of **0.867** on the reserved test set. They used the baseline BART-base with copy mechanism as the generator and leverage adversarial training during fine-tuning by using FGM. They enhance the entities by inserting XML-like tags, and alter the location of constraint and objective direction entities to where they occur in the original input description. For more details, refer to (Ning et al., 2023) or their code[5].

#### 5.2.4. Fourth Place: Team PingAn-zhiniao

Team PingAn-zhiniao primarily leveraged **data preprocessing** and **hyperparameter tuning** to achieve an accuracy of **0.866** on the reserved test set. Data preprocessing included wrapping entity types with tags and they report that the most improvement was brought when the bert_dropout hyperparameter was set to 0.5.

#### 5.2.5. Fifth Place: Team Infrrd AI Lab

Team Infrrd AI Lab **preprocessed** the input and utilized **multitask training** to achieve an accuracy of **0.780** on the reserved test set. They used the text-to-text transfer transformer (T5) (Raffel et al., 2019) and processed the input by wrapping entities with the markup of entity types. They also reported an increase in performance when they separated each sample into multiple samples, each corresponding to one declaration. Multitask learning was leveraged to train the model to generate text when given different prompts. For more details, refer to (He et al., 2022).

### 5.3. Experiments with large language models

After the competition ended, we wanted to compare the performance of black-box large language models. In particular, we conducted some experiments with ChatGPT to see how it would perform in our competition. For these experiments, we combined the two subtasks and directly asked ChatGPT to generate a problem formulation from a given LPWP (problem description). We evaluated the performance of ChatGPT on both the test and

6. Team UIUC code: https://github.com/mlpgroup/nl4opt-eq-generation
7. Team Sjang code: https://github.com/jsh710101/nl4opt

development datasets using the declaration-level mapping accuracy defined in Equation (2). To ensure consistent output from ChatGPT, we structured our prompts as follows:

" *<Problem description> Use the above problem description and write the optimization formulation of the problem. Please only give me the model with just one-line explanations for each model element. I don't need the solution. Remove all non-essential spaces. Don't simplify the expressions and don't use LaTeX code or any code in your responses. Use "x", "y", and "z" as variables name.*"

For these experiments, we used the *gpt-3.5-turbo* model trained on data up to September 1st, 2021. To evaluate the performance of ChatGPT, we asked OR experts to manually verify the correctness of the generated models by ChatGPT and measured the per-declaration accuracy. ChatGPT achieved an accuracy of **0.927** on the reserved test set for this combined task.

## 6. Discussion

**Sub-task 1:** Four of the top 5 teams used ensemble learning to maximize the F1 score for the NER task. While this is a great technique for competitions like NL4Opt that only consider performance metrics, it drastically increases the complexity which makes the training and inference more computationally expensive and less transparent. When considering methods from this competition for real-world time-sensitive applications, methods such as the Student-Teacher learning framework (Wang and Yoon, 2022) could be explored. Other successful techniques included simple augmentation and preprocessing. Some methods also included adversarial training, or training through a two-step approach (i.e., fine-tuning using a global pointer then switch to the multi-head decoder). It is also worth pointing out that many winning teams used the transformer-based language model, DeBERTa, often as part of an ensemble. These winning methods resulted in a 2.3 to 3.3% increase in F1 scores compared to baseline with the highest F1 score of 0.939 by Team Infrrd AI Lab (He et al., 2022).

**Sub-task 2:** The improvements from the winning teams primarily resulted from preprocessing and data augmentation. Every winning team implemented some data augmentation or alterations to the input. The top two submissions replaced BART-base with BART-large which was responsible for higher top accuracy but a higher standard deviation was also reported. This sub-task highlights the importance of the input prompt design. We will continue to explore different input representations and the impact it has on performance. We are also interested in further exploring methods of data augmentation and training methods (i.e., ensemble learning, adversarial learning, etc.). The results of these winning submissions were encouraging as we saw a 17 to 29% increase in declaration-level accuracy from the winning submissions with the highest accuracy of 0.899 by Team UIUC-NLP (Gangwar and Kani, 2022).

**Comparison with large language model** Although ChatGPT was not trained or fine-tuned on our training set, it outperformed the winning submission of sub-task 2 by 2 percentage point. The common errors made by ChatGPT, in order of frequency of occurrence, include incorrect variable coefficients in constraints, extraneous constraints, wrong

constraint directions, extra variables, missing constraints, and incorrect variable coefficients in the objective.

The datasets used in this competition have had a lower level of complexity compared to real-world problems. As a result, it remains unclear how ChatGPT would perform when faced with more realistic and challenging problem descriptions that are frequently encountered in practical scenarios. Therefore, further research is required to examine the generalizability of large language models across a more extensive range of problem descriptions with varying levels of complexity and realism. Furthermore, it is crucial to explore methods for enhancing the trustworthiness and robustness of these models to extend their usefulness in practical applications.

## 7. Conclusion

We hosted the NL4Opt Competition at NeurIPS 2022 to draw attention towards the potential of machine learning in augmenting the user experience of OR tools. This competition presented two engaging tasks that successfully attracted many unique solutions. The two tasks (NER and generation) combine to take a linear programming word problem, tag its relevant entities, and generate a canonical representation that can be easily converted into a format that optimization solvers can interpret.

To summarize, many winning teams of sub-task 1 reported a significant improvement in performance when leveraging ensemble learning and various augmentation techniques. Winning teams of sub-task 2 reported the main contributor for improved performances resulted from redesigning the input prompt. These solutions improved upon the baseline (up to 3.3% for sub-task 1 and 29% for sub-task 2) and will be explored for their use in making commercial solvers more accessible to non-experts by accepting natural language problem descriptions. ChatGPT achieved a 2.8% improvement over the top-performing submission for sub-task 2 without the need for intermediate entity tagging. Future research should investigate the generalizability of large language models and the potential benefits of fine-tuning them for specific applications.

In addition to the impact of providing an alternative input format to solvers, the labelled dataset from this competition has been released and may be used to evaluate methods for multi-sentence inputs, low-resource learning (eg. zero-shot/few-shot learning), and generalizability to unseen domains. We encourage and look forward to continual applications of the open-sourced dataset and the subsequent exciting new research interests that may stem from the solutions of the NL4Opt Competition.

## Acknowledgments

A big thank you to all the participants of the competition for their interest and engagement. This competition was a success thanks to your hard work. We would also like to acknowledge the effort of the co-organizers from University of British Columbia, Ivey Business School, and Huawei Technologies Canada Co. Ltd.

## References

Jeneva Beairsto, Yufan Tian, Linyu Zheng, Qunshan Zhao, and Jinhyun Hong. Identifying locations for new bike-sharing stations in glasgow: an analysis of spatial equity and demand factors. *Annals of GIS*, 0(0):1–16, 2021. doi: 10.1080/19475683.2021.1936172. URL https://doi.org/10.1080/19475683.2021.1936172.

Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1133. URL https://aclanthology.org/P14-1133.

Gabriel R. Bitran and René A. Caldentey. An overview of pricing models for revenue management. *IEEE Engineering Management Review*, 44:134–134, 2016.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2019. URL https://arxiv.org/abs/1911.02116.

Xuan-Dung Doan. Vtcc-nlp at nl4opt competition subtask 1: An ensemble pre-trained language models for named entity recognition, 2022. URL https://arxiv.org/abs/2212.07219.

Li Dong and Mirella Lapata. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1004. URL https://aclanthology.org/P16-1004.

Yujian Gan, Matthew Purver, and John R. Woodward. A review of cross-domain text-to-SQL models. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 108–115, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-srw.16.

Neeraj Gangwar and Nickvash Kani. Highlighting named entities in input for auto-formulation of optimization problems, 2022. URL https://arxiv.org/abs/2212.13201.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014. URL https://arxiv.org/abs/1412.6572.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. Dialog-to-action: Conversational question answering over a large-scale knowledge base. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2946–2955, 2018.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. Benchmarking meaning representations in neural semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1540, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.118. URL https://aclanthology.org/2020.emnlp-main.118.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.38. URL https://aclanthology.org/2021.semeval-1.38.

JiangLong He, Mamatha N, Shiv Vignesh, Deepak Kumar, and Akshay Uppal. Linear programming word problems formulation using ensemblecrf ner labeler and t5 text generator with data augmentations, 2022. URL https://arxiv.org/abs/2212.14657.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. URL https://arxiv.org/abs/2111.09543.

Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2153. URL https://aclanthology.org/S19-2153.

Sanghwan Jang. Tag embedding and well-defined intermediate representation improve auto-formulation of problem description, 2022. URL https://arxiv.org/abs/2212.03575.

N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, page 302–311, New York, NY, USA, 1984. Association for Computing Machinery. ISBN 0897911334. doi: 10.1145/800057.808695. URL https://doi.org/10.1145/800057.808695.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL https://arxiv.org/abs/1910.13461.

Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. NL2Bash: A corpus and semantic parser for natural language interface to the linux operating system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1491.

Yingying Ma, Xiaoran Qin, Jianmin Xu, and Xiangli Zou. Research on pricing method of public bicycle service: A case study in guangzhou. In *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pages 156–161, 2016. doi: 10.1109/SOLI.2016.7551679.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

Shervin Malmasi, Besnik Fetahu, Anjie Fang, Kar Sudipta, and Oleg Rokhlenko. MultiCoNER – multilingual complex named entity recognition. Online, 2022. URL https://multiconer.github.io/competition.

Michael J Maloni and WC Benton. Supply chain partnerships: opportunities for operations research. *European journal of operational research*, 101(3):419–429, 1997.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.92. URL https://aclanthology.org/2020.acl-main.92.

Ines Montani and Matthew Honnibal. Prodigy: A new annotation tool for radically efficient machine teaching. *Prodigy*, 2018. URL https://prodi.gy/docs.

John C Nash. The (dantzig) simplex method for linear programming. *Computing in Science & Engineering*, 2(1):29–31, 2000.

Yuting Ning, Jiayu Liu, Longhu Qin, Tong Xiao, Shangzi Xue, Zhenya Huang, Qi Liu, Enhong Chen, and Jinze Wu. A novel approach for auto-formulation of optimization problems, 2023. URL https://arxiv.org/abs/2302.04643.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL https://arxiv.org/abs/1802.05365.

Yves Pochet and Laurence A Wolsey. *Production planning by mixed integer programming*, volume 149. Springer, 2006.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019. URL https://arxiv.org/abs/1910.10683.

Rindranirina Ramamonjison, Haley Li, Timothy T. Yu, Shiqi He, Vishnu Rengan, Amin Banitalebi-Dehkordi, Zirui Zhou, and Yong Zhang. Augmenting operations research with auto-formulation of optimization models from problem descriptions, 2022. URL https://arxiv.org/abs/2209.15565.

Diana Qing Tao, Martin Pleau, et al. Analytics and Optimization Reduce Sewage Overflows to Protect Community Waterways in Kentucky. *Interfaces*, 50(1):7–20, January 2020. doi: 10.1287/inte.2019.1022. URL https://ideas.repec.org/a/inm/orinte/v50y2020i1p7-20.html.

Kangxu Wang, Ze Chen, and Jiewen Zheng. Opd@nl4opt: An ensemble approach for the ner task of the optimization problem, 2023. URL https://arxiv.org/abs/2301.02459.

Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, jun 2022. doi: 10.1109/tpami.2021.3055564. URL https://doi.org/10.1109%2Ftpami.2021.3055564.

H Paul Williams. *Model building in mathematical programming*. John Wiley & Sons, 2013.

Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification, 2018. URL https://arxiv.org/abs/1809.05679.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017.