DocCGen: Document-based Controlled Code Generation

Sameer Pimparkhede[a]**, Mehant Kammakomati**[b]**, Srikanth G. Tamilselvam**[b]**,
Prince Kumar**[b]**, Ashok Pon Kumar**[b]**, and Pushpak Bhattacharyya**[a]

[a]IIT Bombay
[b]IBM Research
[a]{sameerp,pb}@cse.iitb.ac.in
[b]{mehant.kammakomati2,prince.kumar12}@ibm.com
[b]{srikanth.tamilselvam,ashokponkumar}@in.ibm.com

Abstract

Recent developments show that Large Language Models (LLMs) produce state-of-the-art performance on natural language (NL) to code generation for resource-rich general-purpose languages like C++, Java, and Python. However, their practical usage for structured domain-specific languages (DSLs) such as YAML, JSON is limited due to domain-specific schema, grammar, and customizations generally unseen by LLMs during pre-training. Efforts have been made to mitigate this challenge via in-context learning through relevant examples or by fine-tuning. However, it suffers from problems, such as limited DSL samples and prompt sensitivity but enterprises maintain good documentation of the DSLs. Therefore, we propose DocCGen, a framework that can leverage such rich knowledge by breaking the NL-to-Code generation task for structured code languages into a two-step process. First, it detects the correct libraries using the library documentation that best matches the NL query. Then, it utilizes schema rules extracted from the documentation of these libraries to constrain the decoding. We evaluate our framework for two complex structured languages, Ansible YAML and Bash command, consisting of two settings: Out-of-domain (OOD) and In-domain (ID). Our extensive experiments[1] show that DocCGen consistently improves different-sized language models across all six evaluation metrics, reducing syntactic and semantic errors in structured code.

1 Introduction

The Natural Language to Code (NL-to-Code) task has become pivotal in the intersection of natural language processing and programming. NL-to-Code systems can help engineers write a program efficiently by conveying their intentions at a higher
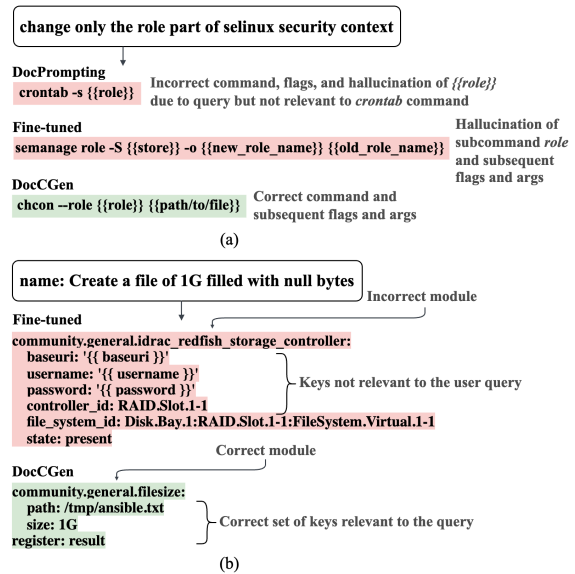

Figure 1: Illustration of shortcomings with fine-tuning and DocPrompting (Zhou et al., 2022) approaches with an example for (a) NL to Bash task (uses GPT Neo 1.3B) and (b) NL to Ansible-YAML task (uses StarCoder2 3B) and the proposed DocCGen method to overcome the limitations.

level, as shown in Figure 1. Systems like Amazon code Whisperer[2], GitHub Co-pilot[3] perform well in NL-to-Code task due to large language models (LLM) trained on extensive data. While they perform well in general resource-rich languages like C++, Python, or Java, their practical usage in structured DSL is limited. DSLs are enterprise-specific languages with specialized schemas and syntax suitable for a specific domain or application[4]. Numerous enterprises use structured languages like Bash, YAML, JSON and HCL (HashiCorp Configuration Language) with specific customizations for automation and to configure and manage infrastructure in IT environments. These languages or their

---

[1]Datasets and code is available at https://github.com/sameerp30/Structured-generation

[2]https://aws.amazon.com/codewhisperer/
[3]https://github.com/features/copilot/
[4]https://w.wiki/6jCH

customizations are potentially unseen by LMs during pre-training, limiting their practical usage (Zan et al., 2022). Some existing methods attempt to address this challenge via in-context learning through examples (Poesia et al., 2022), by fine-tuning (Pujar et al., 2023) or by using relevant documentation as additional context (Zan et al., 2022; Zhou et al., 2022; Parvez et al., 2021; Lu et al., 2022). However, relevant context or samples available for DSL are often insufficient to incorporate diverse library schema rules or specialized structure knowledge in the LM (Zan et al., 2022; Wang et al., 2024). This results in hallucination and different syntactic and semantic errors, as shown in Figure 1. However, enterprises usually maintain detailed documentation of their custom libraries (e.g. ansible modules, bash utilities), including the descriptions, schema, and syntax, to assist developers in enforcing structure and maintaining data integrity. We believe such schema and documentation can be better leveraged during code generation. Therefore, we propose a framework DocCGen that treats the NL-to-Code task as a two-step process, each heavily relying on the documentation. The first step identifies relevant code libraries for the task by retrieving the library documentation relevant to the NL query. The second step employs constrained decoding (CD) to guide code generation by using the grammar and schema rules extracted from the documentation of libraries identified in the first step, as shown in Figure 2. We evaluate this approach for two diverse and complex structured languages, Ansible YAML and Bash command. Generation for these languages is tricky due to complexities like the diverse library schemas, optional and required fields, the order-agnostic nature of fields, and inter-field dependencies. We believe studying these complex structures encompasses most of the challenges in other structured DSLs and allows easily extending DocCGen to other domains. Since the major challenge in DSLs is the limited availability of samples, we focus on enhancing performance for unseen code libraries or libraries with very few samples in the training corpus. Hence, we evaluate our approach in two settings: In-domain and Out-of-domain. Similar to Zhou et al. (2022), none of the libraries in the test set are seen during training in the OOD setting. In the ID setting, every library in the test set has very few NL-to-Code pairs in the train set. DocCGen consistently improve over state-of-the-art models and techniques by a significant

margin (Table 1, 2) across multiple settings.

Finally, we introduce first *publicly* available benchmark dataset for NL to structured code generation task consisting of Ansible-YAML language. Intricate challenges in Ansible-YAML generation, like the complex structure and diverse module schemas, lead to subpar performance even for fine-tuned code LMs (Table 1). We curate `NL to Ansible-YAML` dataset with $18k$ samples with code snippets from more than 2500 modules under OOD and ID settings (Table 5). More information and examples for Ansible-YAML are presented in section A.1. Besides this, we augment new `NL to Ansible-YAML` and existing NL to Bash dataset TLDR (Zhou et al., 2022) with descriptions, detailed schema and grammar information from each library. We believe these datasets will advance research in constrained generation and handling low-resource or unseen data scenarios in structured DSLs.

Our contributions are:

1. Two novel NLP tasks motivated by enterprise needs for precise code generation and validation: a code generation task where models generate code based on fine-grained schema instructions and a code validation task where models validate code samples against given schemas.

2. An extensive study on two diverse structured languages, Bash command and Ansible YAML, for Out-of-domain and In-domain settings. The results show our framework outperforms state-of-the-art techniques across all six metrics (Table 1, 2) for different-sized models.

3. New datasets a) `NL to Ansible-YAML` dataset with $18k$ pairs (refer to Table 5). b) Descriptions and schema of Ansible YAML modules and bash utilities (Section 4) to further motivate research in DSL code generation.

2 Related Work

Constrained decoding: Controlled code generation using constraints has been previously studied majorly for the text-to-SQL task, using plan-based static templates (Bhaskar et al., 2023) or SQL parser-based semantic checks (Scholak et al., 2021). The database schema is fixed and given as input with a text query for text-to-SQL. However, we target a more complex problem involving multiple libraries and diverse schemas and use library
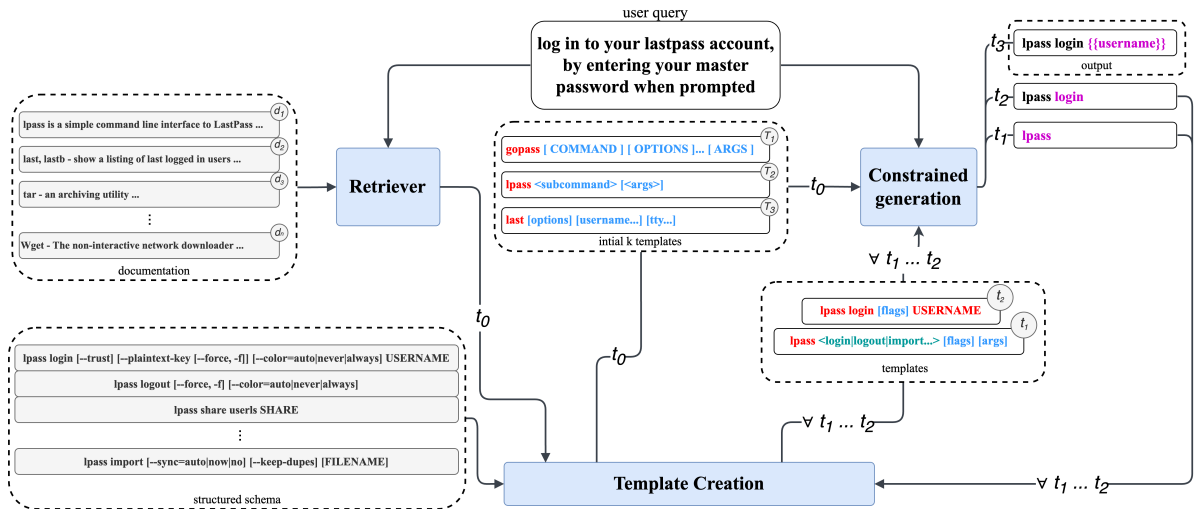
Figure 2: Overview of `DocCGen`. For a given user query, top k relevant library documentations are retrieved and for which initial k templates are created. *Static* part of the template is shown in red, while the *variable* part is in blue. The variable field with a fixed position in the code is enclosed in angle brackets, for instance <subcommand>, as shown in the initial k templates block in the figure. The model is guided to follow one of the templates during decoding. Each time step t_i shows the step-by-step dynamic template evolution and constrained decoding output, adhering to the time-step template leading to the final generated code at $t3$.

documentation to solve this. Poesia et al. (2022) and Wang et al. (2024) use in-context learning via relevant samples or grammar strings and constrain the decoding further. However, in-context learning does not solve the issue of the correctness of the library. Hence, we instead follow a two-step process using library documentation. Agrawal et al. (2023) uses constrained decoding for general-purpose languages like Java and C# using suggestions from intelligent parsers. However, such advanced parsers are uncommon for DSLs and might provide incomplete constraints. Hence, we use rules extracted from documentation more commonly available.

Context Based Controlled Generation like RAGs: Many existing methods retrieve the relevant context and augment it with the input prompt to improve the code generation (Lu et al., 2022; Zan et al., 2022; Zhou et al., 2022; Parvez et al., 2021; Ding et al., 2022). Although effective, these methods do not ensure schema and grammar adherence, especially for unseen libraries and languages. Zhang et al. (2023) and Zan et al. (2022) improve over vanilla retrieval-augmented code generation but require either architectural changes or extra pretraining. Hence, unlike these methods, we guide the generation by adjusting the output logits.

3 DocCGen Framework

`DocCGen` is a two-stage framework: The first stage uses information retrieval (IR) to detect relevant libraries. The second stage uses the neuro-symbolic constrained decoding to control generation and ensure adherence to the schema of relevant libraries.

3.1 Background and Definitions

For a given NL query q, we generate a code snippet c. The first stage of the framework uses a set of documentation D, collected using library descriptions as described in section 4. Hence, each document in D describes the respective library. In this section, we define some frequently used terms.

Structured schema: Structured schema stores the list of valid keywords for every field and the inter-field dependency information. For example, the structured schema of any bash utility (e.g., *cat* or *tar*) includes information like a list of optional and required sub-commands, flags, and inter-field dependency information (e.g., a list of valid flags and arguments for a sub-command).

Template: The template encodes the structure of the code snippet for the library as a string and is used to guide the model during decoding. While the structured schema maintains a list of valid keywords for every field, the template encodes the positional information of fields in the code snippet. Every template has a *static* and *variable* part. The static part is directly copied in the output code, and the model generates the variable part adhering to

18683

the library schema. For Ansible YAML and bash, the template starts with the static part, typically the library name or its variation used in actual code. For example, for the bash utility *git-mv*, template is *git mv [options] {{source}} {{destination}}*. In this template, *[options]* is a variable part and represents the sequence of flags in the command to be generated by the model. The other part is static and is directly included in the output code. Structured schema and template together represent the grammar of the library in the format, which can be easily used to guide the decoding. More example templates are presented in the listing 8.

Trigger signals: Trigger signals G comprises rules to control the generation of optional fields (fields with context-dependent presence and positions) or conditions to dynamically change the template. When triggered, the guiding template changes and makes the model follow new specified rules. For example, generating the " $-$ " token in bash triggers valid doublehand flag generation or generation of pipe operator (token "|") triggers the start of a new process enabling to control generation of command with multiple bash utilities. In YAML, indentation beyond the first level triggers the generation of nested schema with completely different rules from the parent schema, forming a new guiding template. Details of all triggers can be found at A.3.1 and A.1.4.

3.2 Framework

For the given NL query q, the first stage of the framework retrieves k most relevant documents D^* from a pool of documents D. This gives us a set of k most relevant libraries that can be used to generate code c. Then, we fetch the initial templates of every retrieved library stored offline. The next step instantiates the generator model to generate the code snippet c. During auto-regressive inference decoding, the model is constrained to follow one of the k code templates. As the decoding proceeds, the template might be changed dynamically based on the tokens generated by the model, the structured schema of the library, and trigger signals, as shown in Figure 2.

3.3 Information retrieval

We experiment with sparse and dense retrieval systems in the first stage of DocCGen.

3.3.1 Sparse retrieval

We use the BM25 retrieval system (Robertson and Jones, 1976) that uses sparse features such as word frequencies to calculate similarity with documents.

3.3.2 Dense retrieval

For dense retrieval systems, we fine-tune pre-trained ColBERTv2 (Santhanam et al., 2021) and also use it in the zero-shot setting. Finally, we use the best results for the downstream generation task. **Training:** We fine-tune ColBERTv2 based on triplet formed as $< q, D^+, D^- >$. D^+ is the document of the libraries relevant to query q. D^- is a set of documents of libraries that are not relevant to q but are similar to D^+. For q we prepare the training set as $(q, d_1^+, d_2^+, ..., d_m^+, d_1^-, d_2^- ....., d_n^-)$ where d_i^+ is the positive document, and each d_i^- is a negative document which is not relevant to q. We select n *hard negatives* using miniLM sentence BERT similarity scores similar to Santhanam et al. (2021). Using such a train set, we train ColBERTv2 by minimizing the distance between q and D^+ and maximizing the distance between q and D^-.

3.4 Constrained generation

Constrained generation is the second stage of DocCGen. It constrains the model during greedy decoding to follow the library grammar using the template, structured schema, and trigger signals. In this process, if the model has generated (x_1, $x_2, ...x_n$) tokens, x_{n+1} token is sampled from a set of some specific tokens t such that generated code adheres to the library grammar. This is achieved by setting the logits of all tokens outside t to $-\infty$.

This section explains the steps in constrained generation. First, we explain the *string selection* algorithm, which constrains the model to generate a string from a set of strings. This algorithm will be used repeatedly. Constrained generation starts with fetching the initial templates for k retrieved libraries stored offline. Next, *library selection* algorithm constrains the model to adhere to one of the k library templates. As the model adheres to a template, the *generating variable part* algorithm generates value for the variable part of the template as per the library grammar. While generating the variable part, the guiding template might be changed during decoding based on trigger signals and inter-field dependency as explained by *dynamically changing template* algorithm. Finally, required fields are generated as per *generating required fields* algorithm.

String Selection: *String selection* algorithm is used to constrain the model to generate exactly one string from a set of strings (S) $\{s_1, s_2, s_3..., s_n\}$ (Agrawal et al., 2023). Initially, all the strings are tokenized, and we limit the vocabulary V of the model to a set of tokens $t \in V$, which form the prefix of any string in S. Once a token t_i among t is sampled, all the strings that do not have t_i as a prefix are discarded. The same process is repeated until exactly one string is chosen.

Library selection: We traverse all k initial template strings from left to right and collect substrings for each one until the variable part is encountered. As shown in Figure 2, we collect until *gopass*, *lpass*, and *last* as they are static and subsequent parts of text are variable. As soon as the decoding starts, we constrain the model using *string selection* algorithm to generate exactly one of the k substrings. Next, decoding is constrained to follow that template from left to right while adhering to the grammar of the corresponding library.

Generating variable part: Two conditions govern variable part generation. Firstly, when the position and presence of the field are fixed, the model is constrained to select the valid keywords for that using the *string selection* algorithm. Secondly, predefined trigger conditions guide the model in generating from specific string pools when the position or presence varies, determined by query q. For example, the template of the bash command *gh* is *gh <command> <subcommand> [flags]*. In this example, *<command>*, *<subcommand>*, and *[flags]* are the variable parts. The position and presence of *command* and *subcommand* are fixed, and the model is constrained to select the valid keywords for that part using the string selection algorithm. *Flags* is optional, and a pre-defined trigger condition controls its generation.

Dynamically changing template: In many cases, one field's presence depends on another. For example, as shown in Figure 2, the valid flags and arguments change depending on the sub-command generated. Similarly, in Ansible YAML, the rules of the nested schema (optional and required keys) are completely different from those of the parent schema. Hence, if a key with nested schema is produced, the guiding template is changed to follow the rules of nested schema. After generating each variable part, we check field dependency, and if present, we modify the template accordingly.

Generating required fields: The code must include required fields as per schema rules, but their position is not fixed due to the order-agnostic nature of fields. To ensure its presence, we constrain the model to generate the required fields just before the completion of the code. Completion of code is detected by checking for end-of-sequence tokens. This ensures adherence to the schema.

4 Dataset

This section describes datasets for NL to bash and Ansible YAML task, including augmenting datasets with module descriptions and schema information.

4.1 Ansible YAML

We compile the `NL to Ansible-YAML` dataset by extracting data from Google BigQuery and Ansible Galaxy. The dataset comprises over $18k$ of NL to YAML samples, sourced from a diverse collection of more than 2500 modules. We also curate schema rules and descriptions for every module. Schema rules consist of valid optional, required keys and details of the nested schema. We show dataset statistics in Table 5 and more details on data curation in the Appendix A.1.

4.2 Bash command

Since we primarily focus on improving performance for unseen libraries and low-resource data settings, we select the TLDR (Zhou et al., 2022) as our primary dataset for NL to Bash. TLDR consists of 1503 bash utilities across the train and test samples. This data consists of 7342 NL to bash pairs with 4.3 pairs for every utility. Train and test splits of this data consist of 7342 NL to bash pairs. A low number of samples for each utility creates a scarce data scenario.

Other than this, we also use `NL2Bash` (Lin et al., 2018) dataset consisting of 8090 train and 609 test samples for 100 bash utilities. Due to the high number of NL to bash pairs for every bash utility, this dataset allows us to check performance for resource-rich settings. However, Since this is not the major focus of the work, results for `NL2Bash` are included in Appendix (Table 11)

To prepare module descriptions, we use the *description* section of Linux man-pages [5]. Further, we augment the TLDR dataset with the schema rules for each bash utility. Schema information includes a bash command template prepared from *synopsis*

---

[5] https://manned.org/pkg/ubuntu-mantic

section, valid fields (flags and sub-commands), and inter-field dependency information. Schema details and example templates are provided in A.3.

| Model | Bash | | Ansible YAML | |
	Exact Match(%)	Token F1	Schema Correct	Ansible Aware
GPT Neo 1.3B (*)	3.23	31.97	3.11	2.51
GPT Neo 1.3B (+)	4.18	32.78	4.23	3.37
Zhou et al. (2022)	9.05	37.24	-	-
base+IR	5.91	39.20	15.37	10.72
base+IR+CD	**9.40**	**41.26**	**36.58**	**25.19**
StarCoder2 3B (*)	4.09	34.22	4.41	5.80
StarCoder2 3B (+)	3.38	35.53	4.96	5.90
base+IR	7.63	41.67	7.47	4.08
base+IR+CD	**9.56**	**43.25**	**58.82**	**19.76**
StarCoder2 7B (*)	4.12	34.45	5.16	5.61
StarCoder2 7B (+)	5.49	35.72	5.11	5.63
base+IR	8.12	42.12	22.47	11.40
base+IR+CD	**10.21**	**44.09**	**57.00**	**18.37**

Table 1: Results for each fine-tuned language model for OOD setting with and without IR and constrained decoding. Here, the model is constrained to follow the Top-1 retrieved library template only. All the metrics in this table demonstrate the syntactic and semantic correctness of the code. *Model (*)* represents the base fine-tuned model and *model (+)* represents the pre-trained fine-tuned model baseline.

5 Experiments

In this section, we lay out our experiments across NL-to-Code tasks and datasets.

5.1 Experimental settings

We evaluate the performance of our framework on two diverse code languages, Ansible-YAML and bash command. For both tasks, we experiment with two settings involving different train-test splits.
Out of Domain: Here, code libraries in the train and test set are completely disjoint, allowing us to evaluate our method for unseen libraries. We use the original train-test split in TLDR dataset for the bash. For YAML, we randomly split the data into 17647 train and 2056 test samples with 2483 libraries in the train and 365 in the test. OOD split results are demonstrated in Table 1.
In Domain: In this setting, libraries in the test set are a subset of the train set. For bash, we mix the train and test samples of TLDR and re-split them

in the ratio of 85% train and 15% test samples. Further, we filter out the small number of pairs that do not have bash utility in the train set. Finally, we have 6240 train and 1081 test NL to bash command pairs with 1503 unique bash utilities. A similar approach is followed for YAML, which creates 18574 train and 2989 test samples.

5.2 Baselines

Across every task and setting, we establish multiple baselines. The Appendix section A.5.3 describes the hyperparameter details for experiments.

Base (model(*)): Here, we fine-tune the transformer-based decoder-only model for NL-to-Code tasks.

Base + IR: We constrain the base fine-tuned model to follow the template of one of the k retrieved libraries as described by the library selection algorithm (refer to 3.4). However, we do not constrain the model to adhere to its schema for further generation. This allows us to observe the improvement based on the first stage of DocCGen only. Here, we present the results for $k = 1$. Results for $k = 3, 10$ are shown in the Table 7, 8. Further details on pre-training data are provided in the Appendix (section A.2, A.4).

Pre-train (model(+)): Existing methods like APICoder (Zan et al., 2022) pre-train models on abundant documentation and code samples for general-purpose languages like Python. Replicating this setup for structured DSLs is challenging due to the scarcity of available code samples. Hence, for best comparison, we pre-train our models on Linux man pages for bash and Ansible documentation for YAML, ensuring no data leakage from fine-tuning datasets. We then fine-tune the pre-trained model on respective NL-to-Code tasks and compare its performance with DocCGen. We also perform ablation studies with Base + IR setup for the pre-trained models (Table 9, 10). Details of pre-training data are provided in the Appendix (section A.4, A.2).

DocPrompting: We adopt DocPrompting (Zhou et al., 2022) as a baseline for OOD split through the TLDR dataset because it is a RAG-based approach, currently state-of-the-art for TLDR. Additionally, Unlike other RAG-based methods (Parvez et al., 2021; Zhang et al., 2023), it uses documentation instead of abundant code samples, aligning better with our DSL use case with scarce examples.

Model	Bash		Ansible YAML	
	Exact Match(%)	Token F1	Schema Correct	Ansible Aware
GPT Neo 1.3B (*)	8.08	44.02	3.11	2.51
GPT Neo 1.3B (+)	9.12	45.23	4.23	3.37
base+IR	9.12	47.13	15.37	10.72
base+IR+CD	**10.46**	**49.37**	**36.58**	**25.19**
StarCoder2 3B (*)	15.26	50.38	4.65	5.25
StarCoder2 3B (+)	15.26	51.74	4.71	6.20
base+IR	16.31	54.31	6.11	9.22
base+IR+CD	**17.23**	**56.12**	**51.08**	**39.04**
StarCoder2 7B (*)	14.91	50.82	4.38	6.49
StarCoder2 7B (+)	15.63	52.73	4.11	6.39
base+IR	16.79	54.77	7.05	10.43
base+IR+CD	**18.12**	**57.64**	**52.96**	**36.94**

Table 2: Results for each fine-tuned language model for ID setting with and without IR and constrained decoding. Here, the model is constrained to follow the Top-1 retrieved library template only. All the metrics in this table demonstrate the syntactic and semantic correctness of the code.

Model	OOD		ID	
	CMD Acc(%)	Module Match(%)	CMD Acc(%)	Module Match(%)
GPT Neo 1.3B (*)	17.88	18.63	37.01	32.71
GPT Neo 1.3B (+)	17.13	17.01	39.21	33.48
Zhou et al. (2022)	27.59	-	-	-
StarCoder2 3B (*)	17.13	25.12	47.91	52.79
StarCoder2 3B (+)	17.02	26.16	48.38	53.90
StarCoder2 7B (*)	16.16	22.13	46.99	**77.95**
StarCoder2 7B (+)	17.88	21.98	48.38	77.81
+IR/+IR+CD	**38.32**	**36.38**	**60.12**	68.45

Table 3: Results for the library (bash utility or ansible module) detection accuracy in generated code. Here, the model is constrained to follow the Top-1 retrieved library template only. Hence, Command Acc and Module Acc, which detect the exact match of the library in generated code, depend only on IR and give the same scores for IR and IR+CD models.

5.3 Models

Information Retrieval We experiment with sparse retrieval BM25 and dense retrieval ColBERTv2.
Generator We include different sized state-of-the-art code language models in our evaluation, including StarCoder2 family (3B, 7B, 15B) (Lozhkov et al., 2024), and CodeLlama 34B (Roziere et al., 2023). Due to resource constraints to fine-tune large parameter models like CodeLlama 34B

and Starcoder2 15B, we experiment with their instruction-tuned version in a 3-shot setting and present their results in Appendix (Table 6). Further, our evaluation includes a fine-tuned GPT Neo 1.3B (Black et al., 2021) version to compare with the DocPrompting baseline. We use beam search inference decoding for all the base fine-tuned models with beam width 5.

5.4 Evaluation metrics

IR: We evaluate IR using Hits@k metric ($k = \{1, 3, 5\}$). This metric indicates the percentage of accurate documents within the top k retrievals.
Bash command: Evaluation metrics for bash include 1) Command name accuracy (CMD Acc): This metric evaluates the exact match of bash utility in the command (e.g. *tar, cat*). 2) Exact Match: Exact match of full generated command and reference command 3) Token F1 score (Zhou et al., 2022).
Ansible YAML: We leverage 2 evaluation metrics from Pujar et al. (2023) - Schema Correct, and Ansible Aware. Additionally, we introduce the Module Acc metric, which measures the correctness of the generated YAML module. This metric is similar to the CMD Acc metric in bash. Refer to A.1.6 for a detailed description of metrics.

6 Results and Analysis

Results and comparison of our framework with various baselines are presented in Tables 1, 2 and 3. This section presents several observations and a qualitative analysis of the performance.

Improvement in module accuracy: We observe that extended pre-training does not improve performance in structured DSLs with limited code samples in the documentation. Therefore, we use an IR-based approach that focuses on retrieving utility descriptions, unlike Zhou et al. (2022), which retrieves passages with options (flags and sub-commands) and utilities. This targeted detection reduces the search space for IR from $400k$ to $1.5k$ documents, leading to a notable improvement in Hits@1 (Table 4). This improves CMD Acc from 27.59% to 38.32% when the model is constrained to follow the Hits@1 retrieved library template (Table 3). CMD Acc consistently improves for the ID setting by around 6% to 12% (Table 3). For YAML, Module Acc significantly improves compared to the fine-tuned baselines, especially in the OOD setting ($\sim 10\%$). Further, we restrict

	Bash						Ansible YAML					
	Hits@k						Hits@k					
	In Domain			Out of Domain			In Domain			Out of Domain		
	@1	@3	@10	@1	@3	@10	@1	@3	@10	@1	@3	@10
BM25	43.21	56.78	68.34	14.51	21.65	32.57	20.51	30.11	39.78	16.20	24.37	33.12
ColBERTv2 (Zero Shot)	53.43	71.26	78.90	**38.32**	**51.78**	**58.76**	37.69	50.24	61.99	30.30	42.31	55.65
ColBERTv2 (Fine-tuned)	**61.62**	**79.23**	**84.56**	32.21	47.81	54.28	**66.54**	**77.42**	**84.81**	**34.58**	**47.61**	**58.46**

Table 4: Performance of sparse and dense retrieval across NL-to-Code tasks for ID and OOD settings.

the model to follow one of the templates for k retrieved libraries. CMD Acc and Module Acc drop with a higher value of k (Table 7, 8), which is expected since relaxing constraints on the model tend to approach its performance towards the baselines.

Improvement in Code: In the OOD setting (Table 1), fine-tuned code LM baselines struggle to generate correct libraries even for popular languages like Bash, eventually leading to semantically poor code not relevant to the NL query. While, in the ID setting, despite generating correct libraries (indicated by high Module Acc or CMD Acc), baseline models struggle to generate syntactically correct intended code, resulting in subpar Token F1, Schema Correct, and Ansible Aware metric scores (Table 2). This is more pronounced in YAML due to its complex format and diverse schemas. Constraining the model to follow schema rules during decoding restricts the generation of invalid keywords and significantly improves performance across all metrics and settings. For bash, we observe significant improvement (Table 1) over DocPrompting in Token F1 score by leveraging grammar templates from the documentation. For example, for the NL query, *reboot the device from fastboot mode into fastboot mode again*, the ground truth command is shown in Listing 1.

```
# ground truth command
fastboot reboot bootloader

# DocPrompting output command
fastboot reboot path/to/devicefile

# example fastboot command template
fastboot [flags] <flashall|erase
    partition|flashing unlock|reboot
    bootloader|...>

# DocCGen output command
fastboot reboot bootloader
```
Listing 1: Example sample for fastboot command

DocPrompting retrieves correct documents for the given query, which consists of the description of the utility *fastboot* and a document for the subcommand fields *reboot*. Yet it produces an incorrect command as shown in the Listing 1. We instead leverage the template from the *synopsis* and *commands* section of fastboot documentation. As shown in Listing 1, following the grammar template ensures that subcommand is generated from valid strings enclosed in <>. This ensures *reboot* is followed by the word *bootloader*. This approach improves the Token F1 score from 37.24 to 41.26. Hence, constrained decoding using the templates and schema rules reduces the generation of invalid keywords resulting in improved validity of code and agreement with ground truth.

7 Conclusion

We propose DocCGen, a novel framework for NL-to-Code generation for structured DSLs. DocCGen decomposes the NL-to-Code generation into two steps involving the detection of relevant libraries in the first step and using schema and grammar rules extracted from the documentation of these libraries to guide the decoding in the second step. We evaluate the performance of DocCGen for two complex structured languages, Bash command and Ansible YAML, involving two settings, OOD and ID. Our approach outperforms state-of-the-art techniques consistently across all metrics for different-sized models. It reduces syntactic and semantic errors in code, particularly for unseen libraries and low-resource data settings. We also contribute the first *publicly* available benchmark dataset for NL to Ansible-YAML task. We augment NL to Ansible-YAML and TLDR dataset with description and schema information. We hope this work will help advance research in solving DSL-related tasks and constrained generation.

Limitations

We break down code generation in to two steps: a) Information Retrieval and b) Generation based on retrieved documentation. Therefore, errors in retrieval for the user query may cascade to the generation step. Even though, we see that leveraging documentation in this pipeline-based approach results in significant improvements for custom settings, we believe that jointly training the retriever and generator might mitigate these errors. This can be explored as a part of future work. Apart from this, constrained decoding adds a computational overhead during inference. However, since we add the rules on top of efficient greedy decoding, constrained decoding is practical to use as beam search decoding which is widely adopted is similarly computationally heavy. Still, this can be mitigated using constrained generation in speculative decoding similar to Wang et al. (2024). Such improvements can easily be integrated with our framework. Further, parser-based methods to automatically integrate grammar rules during decoding can help generalize DocCGen to a larger scale.

Ethics Statement

Custom curated NL to Ansible-YAML data has been collected from sources like Google BigQuery and Ansible Galaxy, which are publicly available platforms. Other datasets and documents used are from open-source repositories, are publicly available, and can be used without any copyright issues.

References

Lakshya A Agrawal, Aditya Kanade, Navin Goyal, Shuvendu K Lahiri, and Sriram K Rajamani. 2023. Guiding language models of code with global context using monitors. *arXiv preprint arXiv:2306.10763*.

Adithya Bhaskar, Tushar Tomar, Ashutosh Sathe, and Sunita Sarawagi. 2023. Benchmarking and improving text-to-sql generation under ambiguity. *arXiv preprint arXiv:2310.13659*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Yangruibo Ding, Zijian Wang, Wasi Uddin Ahmad, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, and Bing Xiang. 2022. Cocomic: Code completion by jointly modeling in-file and cross-file context. *arXiv preprint arXiv:2212.10007*.

Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D Ernst. 2018. Nl2bash: A corpus and semantic parser for natural language interface to the linux operating system. *arXiv preprint arXiv:1802.08979*.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Shuai Lu, Nan Duan, Hojae Han, Daya Guo, Seungwon Hwang, and Alexey Svyatkovskiy. 2022. Reacc: A retrieval-augmented code completion framework. *arXiv preprint arXiv:2203.07722*.

Md Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. *arXiv preprint arXiv:2108.11601*.

Gabriel Poesia, Oleksandr Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*.

Saurabh Pujar, Luca Buratti, Xiaojie Guo, Nicolas Dupuis, Burn Lewis, Sahil Suneja, Atin Sood, Ganesh Nalawade, Matt Jones, Alessandro Morari, and Ruchir Puri. 2023. Invited: Automated code generation for information technology tasks in yaml through large language models. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–4.

Stephen E. Robertson and Karen Spärck Jones. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27:129–146.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bailin Wang, Zi Wang, Xuezhi Wang, Yuan Cao, Rif A Saurous, and Yoon Kim. 2024. Grammar prompting for domain-specific language generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daoguang Zan, Bei Chen, Zeqi Lin, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2022. When language model meets private library. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 277–288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fengji Zhang, Bei Chen, Yue Zhang, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*.

Shuyan Zhou, Uri Alon, Frank F Xu, Zhengbao JIang, and Graham Neubig. 2022. Doccoder: Generating code by retrieving and reading docs. *arXiv preprint arXiv:2207.05987*.

A   Appendix

We provide additional details for NL to Ansible-YAML, and NL to Bash task, hyperparameter details, and additional analysis on performance in a low resource setting. Firstly we present the details of Ansible-YAML which consists of data collection, schema rules, a list of trigger signals, and evaluation metrics in section A.1. We present the same details for the NL to Bash task in the section A.3. The appendix also consists of results for additional ablation studies like Top-3, Top-10 IR (Table 7, 8) results of in-context learning (Table 6), and ablation studies with pre-training data (Table 9, 10).

A.1   Ansible YAML

YAML is one of the standard code languages used to configure systems declaratively. Ansible is an IT automation tool widely used in enterprises that allows the Infrastructure as Code (IaC) paradigm through Ansible playbooks written in YAML. This section describes examples, data collection, statistics, and evaluation metrics for NL to Ansible-YAML task.

A.1.1   Examples

Some examples (Listing 2 and 3) of Ansible YAML are provided to show glimpse of their syntax.

```
- name: Create a symbolic link
  ansible.builtin.file:
    src: /file/to/link/to
    dest: /path/to/symlink
    owner: foo
    group: foo
    state: link
```

Listing 2: Example Ansible YAML for file module with simple key value pairs

```
- name: Build 'all' target with args
  make:
    chdir: /home/ubuntu/cool-project
    target: all
    params:
      NUM_THREADS: 4
      BACKEND: lapack
```

Listing 3: Example Ansible YAML for make module with nested key value pairs

A.1.2   Data Collection

We curate the dataset from 2 different sources - Google BigQuery and Ansible Galaxy. To curate data from Google BigQuery, we run a SQL query against the BigQuery datastore to pull code files with one of the valid YAML file extensions (.yaml, .yml, .YAML, and .YML). There is no foolproof way to identify Ansible-YAMLs from this corpus. Therefore, we employ simple heuristics based on module keywords and the format of the data to extract Ansible-YAML candidates.

From each Ansible YAML file to subsample NL to YAML candidates, we use a heuristic based on YAMLs having the keys - *name* and *name of the ansible module*. These candidates are then grouped based on the ansible module name and then used for preparing in and out-of-domain settings.

A universal set of Ansible modules is fetched from Ansible Galaxy API along with their documentation. The documentation consists of long and short descriptions, module constraints, and examples. The long and short descriptions are used to prepare data for IR. Examples are combined into NL to Ansible-YAML dataset prepared using Google BigQuery, and module constraints are used in the constrained generation stage.

A.1.3   Data Statistics

Ansible module, NL to Ansible-YAML sample, and YAML key-value pair distribution are shown in Table 5 for both in and out-of-domain settings.

	In Domain		Out of Domain	
	Train	**Test**	**Train**	**Test**
No. of modules	2922	2097	2483	365
No. of samples	18574	2989	17647	2056
Min no. of samples per module	4	1	4	1
Max no. of samples per module	7	7	8	8
Average no. of samples per module	6	1	7	6
Min no. of key value pairs	0	1	0	1
Max no. of key values pairs	1225	97	187	111
Average no. of key value pairs	4	5	4	5

Table 5: Statistics for `NL` to `Ansible-YAML` dataset.

The number of samples per module in both settings does not exceed 8, portraying a low-resource environment.

Some samples have 0 key-value pairs because they are simple strings that still are valid YAMLs. The reason for the total number of modules not being consistent across in-domain and out-of-domain settings is that in the out-of-domain setting for test split, some modules have been dropped as the YAMLs were not valid, and similar data processing has been applied to the in-domain setting as well. Also, the number of modules across the splits for the in-domain setting is not equal because the modules having just 1 sample have been moved to train split to hold the nature of the in-domain setting for the dataset.

A.1.4 Module Description and Structured schema

Ansible Galaxy's API exposes a list of modules and their respective documentation. We use the API to fetch a complete list of modules, and then, for each module, we fetch the module documentation, which includes long and short descriptions. We prepare the module description by appending the short description followed by the long description. We omit those modules which have neither relevant short nor long descriptions. The average length of text descriptions is 816 characters.

We curate schema information from Ansible Galaxy's API, which returns this information as part of the documentation. We augment the dataset with this schema information, which can include valid required and optional keys as shown in Listing 4 and nested schema as shown in Listing 5. Every nested schema further consists of optional and required keys.

```
...
"ise_hostname": {
    "description": [
        "The Identity Services Engine
    hostname."
    ],
    "required": true,
    "type": "str"
},
...
```

Listing 4: Example of type and required key constraints for module device_administration_authentication_rules

```
...
"link": {
    "description": "Device
    Administration Authentication Rules'
    s link.",
    "suboptions": {
        "href": {
            "description": "Device
    Administration Authentication Rules'
    s href.",
            "type": "str"
        },
        "rel": {
            "description": "Device
    Administration Authentication Rules'
    s rel.",
            "type": "str"
        },
        "type": {
            "description": "Device
    Administration Authentication Rules'
    s type.",
            "type": "str"
        }
    },
    "type": "dict"
},
...
```

Listing 5: Example of nested key constraints for module device_administration_authentication_rules

```
# array type
- name: Create a symbolic link
...

# dictionary type
name: Create a symbolic link
...
```

Listing 6: Example prompts for `NL` to `Ansible-YAML` task

Trigger signals: Trigger signals G for YAML are as follows. If the model produces indentation spaces equal to level one keys, it triggers to constrain the model to produce a valid level one

schema by generating valid level 1 keys. Further, if the model generates more spaces, we check the rules for nested schema and constrain the model to adhere to it. If the model generates an invalid indentation, we backtrack, clear the cache of the model, and add the appropriate number of closest indentations in the output. The process of triggering schema rules based on indentation starts to repeat after it.

```
- name: Create a symbolic link
  ansible.builtin.file:
    [force|src|dest|owner|group|state
    ....]: {{gen arg}}

- name: Build 'all' target with args
  make:
    [file|chdir|jobs|make|params|target|
    targets]: {{gen arg}}
```

Listing 7: Example Ansible YAML for file module with simple key value pairs. Here, [a|b|c] denotes one of the values among a,b,c is generated. *gen arg* denotes the argument generated without constraints. The key-value pairs for the next line are controlled again based on indentation generated at the end of the argument.

Enforced schema rules: We ensure that keys generated at every level of YAML adhere to the module schema. YAML consists of optional and required keys. Hence, we ensure that the required keys must be generated in the YAML. We also ensure that none of the keys are duplicated at any level of nesting. The scenario of optional and required keys is followed in the nested schema with keys different than the parent keys. Hence, we follow the rules of nested schema at every level.

A.1.5 Prompt Description

In the case of `NL to Ansible-YAML` task, the prompt is essentially a key-value pair in the YAML, where the key is $name$ and the value is the NL query. The YAML can be an array with one dictionary or a dictionary itself. We show an example in the Listing 6.

A.1.6 Evaluation Metrics

Schema Correct metric evaluates the model on generating schema-compliant YAML, reflecting the YAML's acceptability by the Ansible tool. The Ansible Aware metric captures the closeness of the generated YAML to the ground truth by capturing the coverage of the keys and values in the ground truth. We have not used the Exact Match metric from the original paper as it does not capture the nature of Ansible module keys, which are typically

order agnostic. We introduce Module Acc metric, which evaluates the model's capability to generate the expected module for the given prompt.

A.2 Pre-training data

For ansible pre-training, we append the schema information and descriptions for $2.5k$ modules in a text file [6]. We separate the description and schema information in one document by a newline character and two different ansible documents by two newline characters. We observe that this helps the model better learn the domain knowledge. From every documentation we filter code examples as most of the code examples in the Ansible playbook are present in our custom-curated dataset which we use for fine-tuning. The final pre-training dataset consists of 4.14 million tokens.

A.3 NL to Bash

This section describes specifics of techniques used for NL to Bash task.

A.3.1 Module Description and Constraints

The `TLDR` dataset is not equipped with fine-grained information such as module description and constraints. The dataset has a total of 1503 bash utilities.

Module Descriptions: Document for every bash utility consists of utility descriptions and NL to Bash examples from corresponding bash utility. Details for both components are given below.
Utility Description: We scrape the descriptions of each bash utility from *DESCRIPTION* section of Linux man-pages[7]. Empirically, we observe that the bash utility descriptions are redundant after the first 60 tokens. Therefore, we select the first 60 tokens from the descriptions. However, if the description is shorter than 30 words, we use full documentation as the description.
Examples: For both ID and OOD settings, we augment descriptions of utilities from the train set with two to three NL to bash example pairs. These pairs are randomly sampled from the training corpus itself. For example, if the bash utility *tar* is in the train set, its document is augmented with NL to bash pairs from the train set having utility as *tar*. This ensures that none of the examples from the test set are present in the document. Since utilities
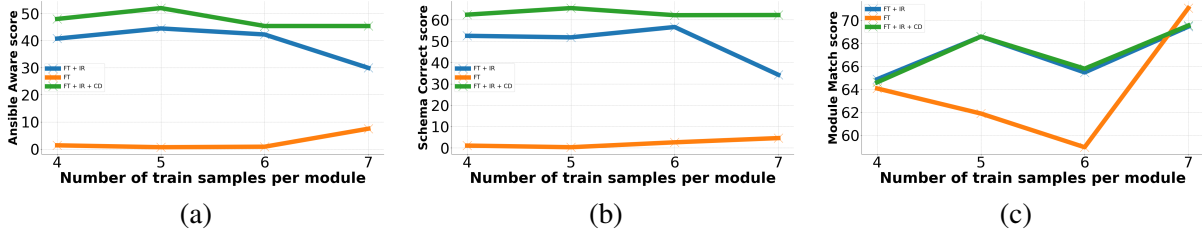

Figure 3: Demonstration of the performance of StarCoder 1B for `NL to Ansible-YAML` task over varying number of train samples per module for in domain setting.
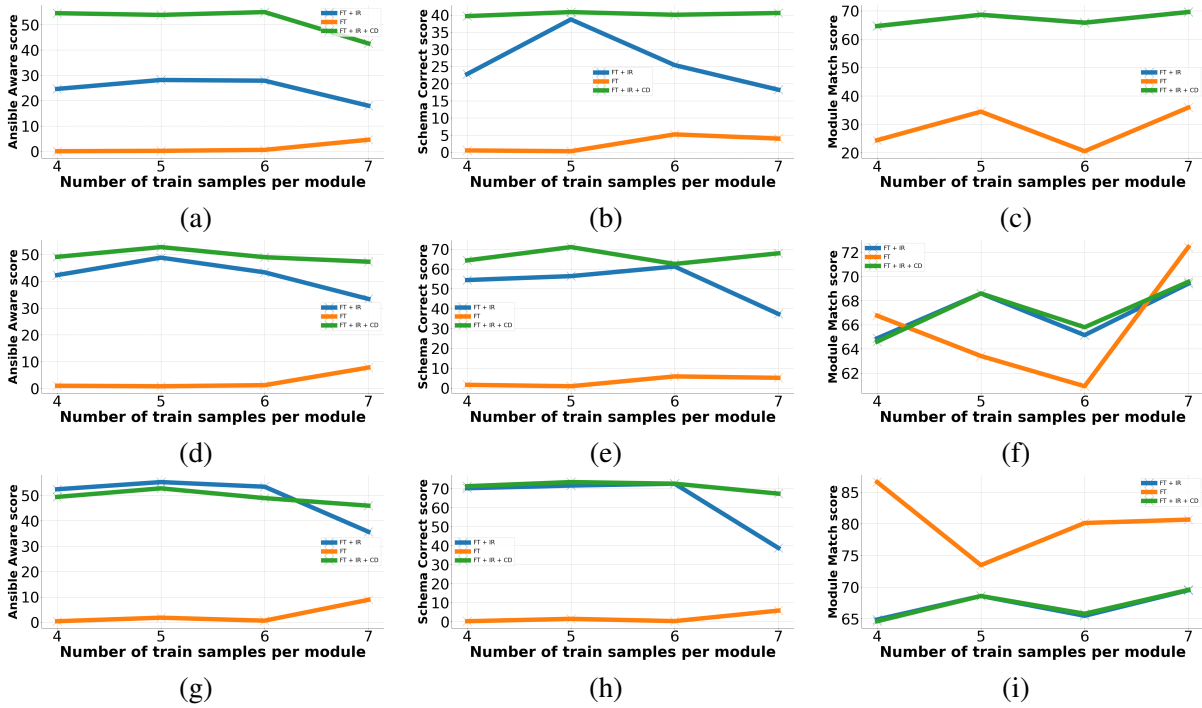

Figure 4: Demonstration of the performance of (a) (b) (c) GPT Neo 1.3B, (d) (e) (f) StarCoder2 3B, and (g) (h) (i) StarCoder2 7B in different configurations for `NL to Ansible-YAML` task over varying number of train samples per module for in domain setting. We omit CodeLlama 34B as it is evaluated in few-shot setting.

in the OOD split test set are disjointed from the train set, documents for the utilities in the OOD split test set consist of only utility descriptions.

```
cp [OPTION] {{SOURCE}} {{DIRECTORY}}

needrestart [-{{v|q}} | -n | -c <cfg> |
    -r <mode> | -f <fe> | -u <ui> | -{{b
    |p}} | -kl]

git rename-tag {{old-tag-name}} {{new-
    tag-name}}

lzop [ command ] [ options ] [ filename
    ... ]

meson setup [ options ] [ build
    directory ] [ source directory ]
```

```
gh <command> <subcommand> [flags]
```

Listing 8: Example templates for bash command curated using synopsis section in linux man page. Here fields within [] denotes optional fields and [a|b|c] denotes that one of the strings among from a, b or c has to be generated

Structured schema: We augment TLDR dataset with schema information for every bash utility. We crawl the Linux man pages of bash modules and collect the initial template T of the bash command for each library from *usage* or *SYNOPSIS* section. Further, we collect the list of valid options and sub-commands for each bash utility. Schema information also includes inter-field dependency information, like a list of valid flags and arguments for every subcommand. For example, for the Linux command *cp*, some of the valid options are *-a, −*

Model	Bash			Ansible YAML		
	Exact Match (%)	CMD Acc (%)	Token F1	Module Acc (%)	Schema Correct	Ansible Aware
Codellama 34B (3 shot)	13.2	32.4	21.8	12.35	20.33	3.54
+ IR	16.71	**38.32**	26.49	**36.38**	13.18	7.39
+ IR + CD	**19.63**	**38.32**	**29.71**	**36.38**	**65.72**	**15.77**
StarCoder2 15B (3 shot)	11.78	30.71	19.63	11.06	4.32	0.53
+ IR	15.62	**38.32**	24.71	36.38	12.05	3.40
+ IR + CD	**18.19**	**38.32**	**31.83**	**36.38**	**66.04**	**20.78**

Table 6: Results for in-context learning for out-of-domain setting with and without IR and constrained decoding. Here, the model is constrained to follow the Top-1 retrieved library template only. Hence, Command Acc and Module Acc, which detect the exact match of the library in generated code, depend only on IR and give the same scores for IR and IR+CD models.

Model	Bash			Ansible YAML		
	Exact Match (%)	CMD Acc (%)	Token F1	Module Acc (%)	Schema Correct	Ansible Aware
StarCoder2 3B	4.09	17.88	34.22	25.12	4.65	5.35
+ IR (Top 3) + CD	**5.24**	**27.33**	**36.50**	**27.29**	**49.45**	**17.66**
+ IR (Top 10) + CD	4.88	25.31	34.91	24.52	47.8	15.25
StarCoder2 7B	4.12	16.16	34.45	22.13	5.16	5.61
+ IR (Top 3) + CD	**5.61**	**26.41**	**37.71**	**25.41**	**47.81**	**19.32**
+ IR (Top 10)+ CD	4.31	24.14	33.73	23.82	45.62	17.14

Table 7: Results for each base fine-tuned language model for out-of-domain setting with and without IR (top 3 and 10 retrievals) and constrained decoding.

archive, *-f, –force*, and *-i, –interactive* are scraped from linux man page.

Templates: Along with options, we also scrape the syntax of bash modules mentioned under *usage* section. In *SYNOPSIS* section, it is standard practice that text enclosed within [] is optional, and the presence and position of that field in the command are not fixed. Text enclosed within <> must be produced at the position in the template. For the optional fields, we use language-specific trigger signals G. Examples of bash command templates are given in listing 8.

Trigger signals: Trigger signals used for bash are as follows. If the model generates the token " –," we constrain the model from generating the string from valid doublehand flags. Similar constraints are used for shorthand flags " -". Other trigger signals include the generation of a pipe operator ("|"). In the bash command, the pipe operator forwards the output of one process to another as input. For example, bash command *nl -s prefix file.txt | cut -c7-* consists of two bash utilities *nl* and *cut*

separated by "|". Generation of token "|" denotes the start of a new process with a new bash utility. Hence, while decoding, if the model generates an operator-like token ("|"), then we constrain the model to freshly follow one of the k templates from the start using the library selection algorithm again 3.4. This trigger signal allows us to generate the bash command with multiple utilities or processes.

Enforced schema rules: We ensure that all the required fields (flags and subcommands) are generated according to their position specified in the template. Further, it is also ensured that all the generated flags and subcommands adhere to the library schema. For the templates that specify the compulsory arguments, we treat those arguments as *static* part of the template and include it in the final output. For example, as given in the template of bash utility *cp*, *source* and *directory* are the compulsory arguments and hence directly included in the output command.

Model	Bash			Ansible YAML		
	Exact Match (%)	CMD Acc (%)	Token F1	Module Acc (%)	Schema Correct	Ansible Aware
StarCoder2 3B	15.26	47.91	50.38	52.79	4.65	5.25
+ IR (Top 3) + CD	**16.71**	**54.55**	**54.31**	**56.21**	**49.37**	**36.21**
+ IR (Top 10) + CD	15.51	53.22	52.89	46.62	47.56	34.24
StarCoder2 7B	14.91	46.99	50.82	**77.95**	4.38	6.49
+ IR (Top 3) + CD	**16.27**	**53.44**	**54.07**	58.56	**47.13**	**33.51**
+ IR (Top 10)+ CD	15.22	51.15	52.49	50.15	45.38	30.76

Table 8: Results for each base fine-tuned language model for in-domain setting with and without IR (top 3 and 10) retrievals) and constrained decoding.

Model	Bash			Ansible YAML		
	Exact Match (%)	CMD Acc (%)	Token F1	Module Acc (%)	Schema Correct	Ansible Aware
StarCoder2 3B	4.18	17.13	32.78	26.16	4.96	5.90
+ IR (Top 1)	5.12	**38.32**	39.81	36.38	22.47	11.12
+ IR + CD	**6.24**	**38.32**	**41.73**	**36.38**	31.21	16.26
StarCoder2 7B	5.49	17.88	35.72	21.98	5.11	5.63
+ IR (Top 1)	**6.23**	**38.32**	40.71	36.38	3.93	3.23
+ IR + CD	7.81	**38.32**	42.31	36.38	43.43	16.38

Table 9: Results for each pre-trained and further fine-tuned language model for OOD setting with and without IR (top 1) and constrained decoding.

A.4 Pre-training data

We append the Linux man-pages for $1.5k$ bash utilities in a single file which is used for pre-training[8]. For every man page, we remove all newline characters and replace double newline characters with a single newline. This keeps the definition of each flag and field separate from each other and results in better performance. The final pre-training data consists of 10.3 million tokens.

A.5 Hyperparameter details

We use NVIDIA A100 80 GB GPUs to perform inference and training for all the experiments. We use the standard HuggingFace transformers (Wolf et al., 2020) with accelerate to load, train, and perform inference for all the models. For constrained decoding we use HuggingFace logits processor[9].

A.5.1 Ansible YAML

All fine-tuned models are fully parameter-tuned to the task. For fine-tuning, we used Adam optimizer with batch size two for all the models and context

length of 2048. We also use the linear learning scheduler and a learning rate of $4e-5$. At inference, we experimented with both greedy search and beam search-based decoding techniques for baselines, and we observed beam search with 5 number of beams performed the best. Training is done for two epochs. All the models are used in bf16 precision.

We use the bert-based-uncased model as base and fine-tune the standard ColBERTv2 pre-trained model[10] on NL to Ansible-YAML task. The document corpus size is 2922 documents. We run the fine-tuning task for 5000 max number of steps. We use 8 negatives for every query while preparing the triplets. The train-test splits for fine-tuning follow the numbers from language model fine-tuning (Table 5).

A.5.2 Bash command

All the training details for bash command generation are the same as those for ansible YAML, except that we use a batch size of 4 with gradient accumulation steps of 4 during fine-tuning. The maximum sequence length for the bash command is 512. All the models are used here in fp32 precision.

---

[8] https://manned.org/
[9] https://huggingface.co/docs/transformers/en/internal/generation_utils#transformers.LogitsProcessor

[10] https://github.com/stanford-futuredata/ColBERT

Model	Bash			Ansible YAML		
	Exact Match (%)	CMD Acc (%)	Token F1	Module Acc (%)	Schema Correct	Ansible Aware
StarCoder2 3B	15.26	48.38	51.74	53.90	4.71	6.20
+ IR (Top 1)	16.71	**60.12**	54.61	68.45	39.11	35.41
+ IR + CD	**17.81**	**60.12**	**56.73**	**68.45**	**48.41**	**38.98**
StarCoder2 7B	15.63	48.38	52.73	**77.81**	4.1	6.39
+ IR (Top 1)	**16.21**	**60.12**	**54.77**	68.45	45.60	40.61
+ IR + CD	15.22	60.12	52.49	68.45	**52.09**	**42.66**

Table 10: Results for each pre-trained and further fine-tuned language model for in-domain setting with and without IR (top 1) and constrained decoding.

Similar to `NL to Ansible-YAML` task, we use the pre-trained ColBERTv2 for fine-tuning the task data. The document corpus size is 1503 documents. Similar to `NL to Ansible-YAML` task, we run for a max of 5000 number of steps. We use 8 negatives for every query while preparing the triplets.

A.5.3 Pre-training

For pre-train the language models on the next word prediction task using library documentation for 3 epochs. For pre-training we use a cosine scheduler with a learning rate of $5e-05$. We experiment with both linear and cosine schedulers and use cosine scheduler checkpoints for further fine-tuning due to the best results. We pre-train with a batch size of 4, gradient accumulation steps of 8, and bf16 precision. Due to scarce data, we use warmup steps of 100 for bash and 150 for ansible pre-training. We use the block size of 1024 for pre-training.

A.6 Analysis

Promising low data resource performance: First, DocCGen outperforms all the baselines in the OOD setting (Table 1) and performs competitively across overall degrees of low-resource data (Figure 3) in ID setting. Second, the performance of fine-tuned StarCoder2 3B in generating good YAML code following the ansible module improves gradually for Ansible Aware and Schema Correct metrics with an increase in training samples. However, extrapolating this growth to meet DocCGen's performance might require a large number of training samples per module. Third, DocCGen outperforms baselines in most of the lower orders of training sample count for Module Acc metric. This behavior is consistent across all models. (Figure 4)

Model	Bash		
	Template Match (%)	Command Acc (%)	Token F1
StarCoder 1B	14.32	57.34	58.42
+ IR + CD	**18.92**	**73.24**	**66.47**
StarCoder 3B	16.34	61.34	62.34
+ IR + CD	**18.39**	**73.87**	**66.89**

Table 11: Results for NL2bash dataset using Top-1 IR