

# AutoDSL: Automated domain-specific language design for structural representation of procedures with constraints

Yu-Zhe Shi\*, Haofei Hou\*, Zhangqian Bi, Fanxu Meng, Xiang Wei,  
Lecheng Ruan<sup>✉</sup>, Qining Wang<sup>✉</sup>

Department of Advanced Manufacturing and Robotics,  
College of Engineering, Peking University

\*Equal contribution ✉ {ruanlecheng, qiningwang}@pku.edu.cn

## Abstract

Accurate representation of procedures in restricted scenarios, such as non-standardized scientific experiments, requires precise depiction of constraints. Unfortunately, Domain-specific Language (DSL), as an effective tool to express constraints structurally, often requires case-by-case hand-crafting, necessitating customized, labor-intensive efforts. To overcome this challenge, we introduce the AutoDSL framework to automate DSL-based constraint design across various domains. Utilizing domain specified experimental protocol corpora, AutoDSL optimizes syntactic constraints and abstracts semantic constraints. Quantitative and qualitative analyses of the DSLs designed by AutoDSL across five distinct domains highlight its potential as an auxiliary module for language models, aiming to improve procedural planning and execution.

## 1 Introduction

Comprehending and executing procedures articulated in natural language to achieve a specified goal represents a fundamental challenge for Artificial Intelligence (AI) systems. With the boost of Large Language Model (LLM) (Brown et al., 2020; Raffel et al., 2020; Touvron et al., 2023), AI systems possess the capability of reasoning over and planning for procedural tasks intended for both human and robotic execution across a broad spectrum of everyday scenarios<sup>1</sup> (Pareti et al., 2014a,b; Tandon et al., 2020; Donatelli et al., 2021; Zhou et al., 2022b), such as cooking according to a recipe, obviating the necessity for external representation of procedures beyond text.

However, contexts significantly more restricted than everyday scenarios, such as conducting non-standardized experiments in scientific laboratories, need to follow specific *protocols*<sup>2</sup>. A proto-

<sup>1</sup>Visit [www.wikiHow.com](http://www.wikiHow.com) for demonstrations.

<sup>2</sup>Visit [www.nature.com/nprot/](http://www.nature.com/nprot/) for examples.

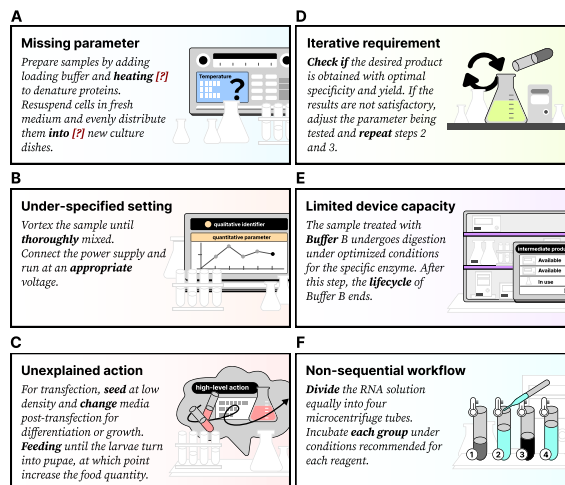


Figure 1: **Representative constraints in protocols.** (A) Parameter omission: This refers to the absence of essential parameter values within a predefined set, *e.g.*, the lack of temperature specification during the denaturation step in Protein Gel Electrophoresis. (B) Parameter under-specification: This occurs when a quantitative parameter is described using qualitative terms, leading to ambiguity, *e.g.*, unclear mixture configurations in DNA Extraction. (C) Action undefinition: This involves the description of procedural steps at a high level without grounding to the specific, executable actions required, *e.g.*, the vague *change* operation in Cell Preparation. (D) Iterative control logic: Loops that operate iteratively to satisfy a final condition rather than straightforwardly, as seen in PCR Optimization. (E) Memory management: Drawing a parallel with computer memory mechanisms, laboratory procedures also face constraints on the availability of storage for intermediates, necessitating explicit reallocation of containers and devices, such as managing buffers in Protein Digestion. (F) Concurrent management: The synchronization of actions without dependencies to maximize time efficiency and resource utilization, *e.g.*, reagent splitting in RNA Extraction.

col delineates every aspect of an experiment’s procedure to facilitate its reproduction (Baker, 2016; Munafò et al., 2017), emphasizing the necessity for precision in every step, to ensure accurate execution by an experimenter. The complexity of procedures, limitation in resources, and susceptibility to error in such scenarios render any deviation from the established protocols inadmissible. Unfortunately, natural language inherently possesses ambiguities (Russell, 1923). Within protocols, description of actions can be semantically

| Protocol text   | Resulting DSL  | BioCoder DSL   | Knowledge Graph  |
|---|--|--|--|
| Repeat the DNA purification process three times using fresh reagents and buffers for each cycle, following the specified centrifugation conditions at 12,000-16,000 g for 1-2 minutes in a microcentrifuge. | <pre>repeat (3){   purify(     reagent="DNA",     device="microcentrifuge",     force=range(       12000g,       16000g),     duration=range(       1min,       2min)) }</pre> | <pre>for (int i=0; i&lt;3; i++){   purify_DNA(     "DNA",     "microcentri fuge",     "refine the DNA's purity" ); }</pre> | <pre>[   ["repeat",   "3"],   ["action",   "purify"],   ["REG",   "DNA"],   ["Device",   "microcentri fuge"] ]</pre> |

Figure 2: **Protocols in different structural representations**

under-specified, and the logic of procedure may be non-linear, as demonstrated in Fig. 1. Given these unique distinctions against daily procedures, accurate interpretation of protocols requires explicit depiction of constraints (see Fig. 2 for example).

Intuitively, structural representation imposes constraints on the processing of protocols. This is achieved through purely symbolic approaches that depict procedures as flow-graphs (Momouchi, 1980; Zhang et al., 2012; Kiddon et al., 2015), and neuro-symbolic hybrid methods that superimpose procedural structures onto neural networks (Puig et al., 2018; Huang et al., 2021, 2022; Zhou et al., 2022a; Zhang et al., 2023; Brohan et al., 2023). Both strategies impose constraints on the interpretation of procedures, thereby reducing the incidence of superfluous operations. Symbolic constraints range from elementary grammars to general-purposed programming languages, with the capability of a constraint system being pivotal in refining a coarse interpretation space into a more precise one (Tarski, 1946; Chomsky, 1957; Hopcroft et al., 1996; Russell and Norvig, 2010).

In light of the context, what level of capability should we expect the constraint to possess? This discourse introduces DSL, a category of symbolic systems endowed with the most potent constraints. DSLs are programming languages tailored to specific problem domains, encapsulating both **syntactic constraints** and **semantic constraints** inherent to those domains. For instance, BioCoder (Ananthanarayanan and Thies, 2010), developed by Microsoft, is a DSL explicitly designed to constrain experimental protocols. On the syntactic level, the variable management mechanism inherited from C/C++ enables DSLs to monitor the lifecycle of each intermediate product, ensuring no omissions or duplications. On the semantic level, the precise definitions of actions, combinations of reagents' names and volumes, with subprocedures abstracted from domain-specific concepts, guaranteeing procedural execution consistency.

Drawing inspiration from DSLs, can we design constraints for protocols in DSL fashion? Hardly, due to the deterministic and substan-

tial cost. Structural constraints necessitate custom design for particular domains, which is prohibitively expensive, given that these domains are highly specialized and often diverge significantly from the conventional purview of computer scientists (Mernik et al., 2005; Fowler, 2010). The development of a DSL necessitates a comprehensive integration of in-depth domain knowledge. Furthermore, the designed DSL must align with theoretical aspects of formal language design while also meeting the distinct requirements of the specialized domain. This necessitates a bidirectional alignment between computer scientists and domain experts, a process that is intrinsically case-by-case, implying that a DSL developed for one domain is unlikely to be applicable or easily adaptable to another, thus limiting the scalability of DSL-based constraints across various domains.

In this study, our objective is to offer an initial proof-of-concept aimed at reducing the design cost of DSL-based constraints for protocols. We propose a scalable framework, termed AutoDSL, that facilitates the automated creation of DSLs. The framework approaches the task as a bidirectional optimization problem, where the design of a DSL is abstracted from domain-specific corpora through a bottom-up process and concurrently derived by general programming language design principles in a top-down manner. This approach emulates the iterative dialogue between computer scientists and domain experts, progressively bridging the conceptual gap between their respective fields of expertise. The syntactic constraint should adequately define consecutive actions and their repetitions, interruption, concurrence, subcomponents, and reactants. Constructs of the semantics constraint need to accurately reflect the domain's concepts and the relations between them, without redundancy or incompleteness.

We utilize protocols from various domains within the experimental sciences — namely *Genetics, Medical and Clinical Research, Ecology and Environmental Research, Bioengineering, and Computational Biology* — as the primary testing ground for our methodology, due to their inherent complexity, resource constraints, and susceptibility to errors. These domains exhibit significant disparities both in syntactic and semantic language features. Comprehensive experiments demonstrate that AutoDSL is capable of generalizing DSL-based constraints tailored to these di-

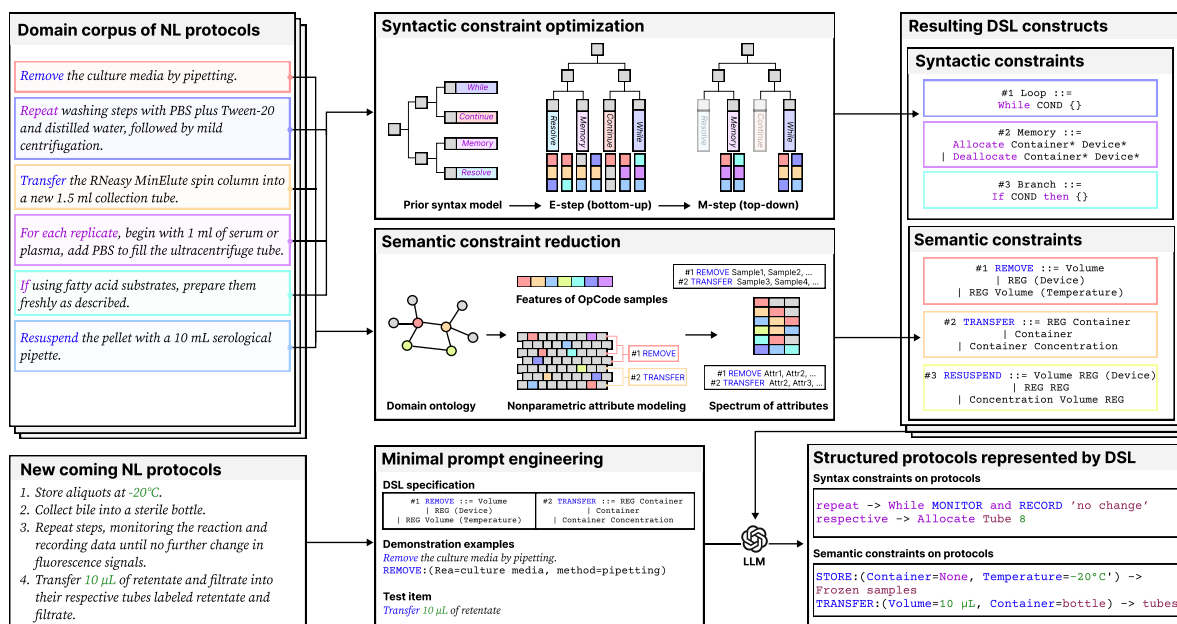


Figure 3: **The AutoDSL framework and the resulting DSL-based procedure constraints.** (Top) Given domain specified corpus, AutoDSL conducts bidirectional syntax optimization and non-parametric semantic reduction, resulting in syntactic constraints and semantic constraints. (Bottom) A DSL-based constraint takes novel procedures as input, handles the nonlinear syntax structures like `loop` by syntactic constraints, and handles the semantic errors like `missing` by semantic constraints.

verse domains, upholding the integrity from both programming language design and domain expertise perspectives. We further demonstrate that syntactic and semantic constraints effectively work as an auxiliary module of LLMs in the processing of unseen protocols, thereby suggesting a promising future for constraining protocols through a synergistic blend of programs and natural language.

The contributions of this work are threefold: (i) We introduce the AutoDSL framework for automated design of DSL-based constraints, which includes a bidirectional syntax optimization module and a non-parametric learning module for semantic reduction. (ii) We establish a systematic and end-to-end evaluation platform to assess the quality of the designed DSLs-based constraints, employing both quantitative and qualitative metrics. (iii) We showcase the efficacy of DSL-based constraints in processing new coming protocols with syntactic complexity and semantic errors.

## 2 Constraints in protocols

In this section, we scrutinize the requirements for precisely constraining protocols. The distinctive challenges of engaging with such systems stem from their complexity of procedures, limited in resources, and vulnerability to errors.

**Complexity of procedures** The complexity of protocols arises from the multitude of action types necessary for conducting experiments, the

extensive categories of reagents involved, the variety of containers and devices for operational implementation, and the broad array of additional conditions affecting action execution, such as duration, temperature, volume, lighting, and acidity. For instance, whereas each daily procedure on wikiHow involves about 14 steps, 27 actions, and 45 objects on average<sup>3</sup>, this figure substantially increases in the context of the five experimental science domains, where each protocol possesses around 60 steps, 76 actions, and 180 objects on average, with an increment of 250% to 350%. This necessitates the utilization of more specialized data types — such as Operations, Reagents, Conditions — over general data types like integers, floats, characters, and strings. Consequently, procedural workflows may include nonlinear elements like loops for repetition, branches for parallel options, and subprocedures for nested and reusable actions, requiring specialized control flow structures to depict these complex scenarios.

**Limitation in resources** The execution of protocols is limited by the availability of resources, including a finite stock of reagents, a limited number of containers, and a scarcity of critical devices, characteristic of medium-scale experimental science laboratories. This parallels the allocation of registers and memory in computing,

<sup>3</sup>The statistics are calculated from the WikiHow Dataset, with 230k procedures in total (Koupaee and Wang, 2018).

where each action must account for the use of memory spaces. In experimental procedure execution, if a device is currently in use, any operations requiring that device must be deferred until it becomes available. This necessitates resource management strategies similar to those in computer programming, such as allocating resources for use and deallocating them post-use. When allocated, the resource becomes inaccessible, creating bottlenecks that introduce inefficiency. To mitigate this, operations that do not compete for the same resources may be executed in parallel.

**Vulnerability to errors** Execution of protocols cannot tolerate errors, mandating strict adherence to every operation. This makes the system exceptionally vulnerable, in contrast to the robustness of everyday procedural execution. In such a context, protocol execution should encompass: (i) syntactic constraints specifying how actions are sequenced to form valid execution; and (ii) semantic constraints verifying that the reactants and reaction conditions are correctly utilized.

### 3 AutoDSL framework

This section describes our problem formulation and solutions for the AutoDSL framework.

#### 3.1 Problem formulation

**Input** The system input consists of natural language descriptions of protocols, namely a domain-specific corpus. These descriptions encompass procedural knowledge of execution orders and ingredient knowledge of reagents, containers, and devices involved in the experiments. For a specific domain, the input corpus  $\mathcal{C} = \{c_1, c_2, \dots, c_N\}$  includes  $N$  protocols.

**Output** The desired output is a DSL  $\{\mathcal{S}, \Lambda\}$ , incorporating sets of constructs that define both syntactic and semantic constraints specific to the domain (Fowler, 2010), while retaining the abstract qualities of a programming language. The set  $\mathcal{S} = \{\varphi_1, \dots, \varphi_K\}$  comprises  $K$  atomic syntactic constraints with production rules, such as control structures `Loop`, `Parallel`, `Jump`, and `Split`. The set  $\Lambda = \{t_1, \dots, t_L\}$  embodies  $L$  atomic semantic constraints, such as operations `Add`, `Remove`, `Incubate`, and `Place`.

#### 3.2 Syntactic constraint optimization

**Key insight** Leveraging existing knowledge on programming language design, our method

utilizes a bidirectional optimization strategy to formulate the syntax of the target DSLs based on the prerequisites of the domain corpora. The algorithm employs an Expectation-Maximization (EM) framework, where the E-Step abstracts syntax from domain corpora and the M-Step derives syntax from programming language principles.

**Modeling** The algorithm models latent syntactic constraint assignments  $\mathcal{Z} = \{z_1, \dots, z_N\}$  for each protocol  $c_i$ . A filter set  $\Theta = \{\theta_1, \dots, \theta_{K'}\}$ , where  $K' > K$ , is designed to determine if a segment of procedural text aligns with the logic of any atomic syntactic constraint, coming with the belief function  $p(\Theta|\mathcal{S})$ . The observational likelihood is computed as  $p(\mathcal{C}|\mathcal{Z}, \Theta) = \prod_{i=1}^N p(x_i|z_i, \theta_{z_i})$ . Hence, the overall joint probability of the model is given by:

$$p(\mathcal{C}, \mathcal{Z}, \Theta|\mathcal{S}) = p(\mathcal{C}|\mathcal{Z}, \Theta)p(\mathcal{Z}|\mathcal{S})p(\Theta|\mathcal{S}). \quad (1)$$

**Syntax prior** Programming language designers leverage a general set of syntactic production rules as the prior  $p(\mathcal{Z}|\mathcal{S})$  for syntax specification. We initialize the set  $\mathcal{S}_0$  with a Context-free Grammar (CFG) (Hopcroft et al., 1996) (see Fig. 4A). Furthermore, we construct the prior belief function  $p(\Theta|\mathcal{S})$  with a series of sliding-window-based filters  $f : \mathcal{C} \mapsto \mathbb{R}$ , which gives a relaxed lower bound for predicting the existence of an atomic syntactic constraint. Please refer to Appx. C for details.

**E-Step** In each E-Step, we obtain the posterior of latent variables  $p(\mathcal{Z}|\mathcal{C}, \Theta, \mathcal{S})$  applying Bayes' theorem, which is implemented by scanning the filters over domain corpus. To note, as the spaces of prior and observation are not intractably large, we simply employ the naive version of E-Step without variational approximations.

**M-Step** In each M-Step, we first maximize the atomic syntactic constraints  $\mathcal{S}$  by maximizing:

$$\mathcal{Q}(\hat{\Theta}, \Theta) = \mathbb{E}_{\mathcal{Z}|\mathcal{C}, \Theta} [\log p(\mathcal{C}, \mathcal{Z}, \hat{\Theta}|\mathcal{S})], \quad (2)$$

where  $\hat{\Theta}$  is the updated  $\Theta$ , resulting in the structural change of  $\mathcal{S}_0$ . These two steps alternate iteratively until convergence (see Fig. 4C), ensuring the syntactic constraints are aligned with the domain.

#### 3.3 Semantic constraint reduction

**Key insight** Following the adaptation of syntactic constraints to the target domain, the semantic reduction phase focuses on distilling fine-grained semantic constraints. This stage addresses



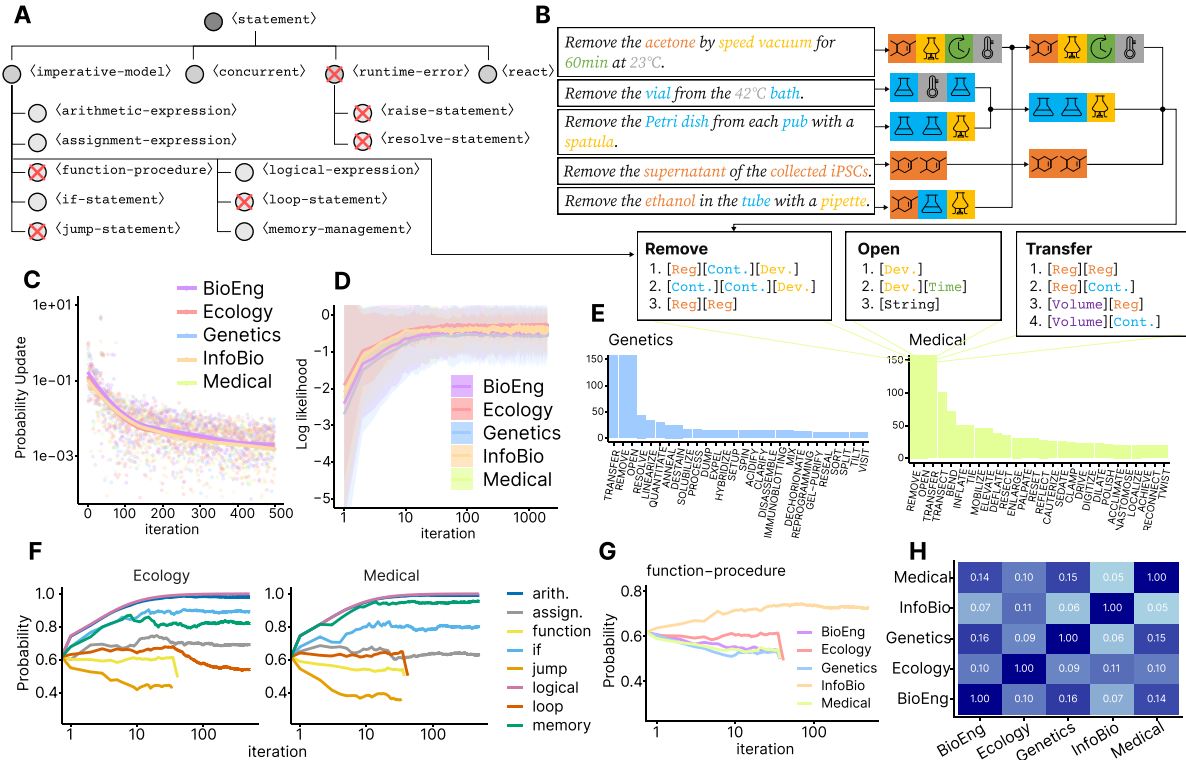


Figure 4: **Illustration on syntactic constraint optimization and semantics constraint reduction.** (A) Resulting syntactic constraints derived from the CFG prior model. (B) Resulting semantic constraints. (C) Convergence curve of syntactic constraint optimization. (D) Convergence curve of semantic constraint reduction. (E) Frequency profile of the semantic constraints of Genetics-DSL and Medical-DSL. (F) Acquisition of different syntactic constraints on Ecology and Medical domain corpora. (G) Syntactic constraint `Function-procedure` is differently acquired by the five distinct domains. (H) Confusion matrix indicating the overlapped semantic constraints between the five distinct domains.

the absence of the domain-specific prior knowledge on semantics by employing a non-parametric approach, *i.e.*, Dirichlet Process Mixture Model (DPMM), which allows for the flexible identification of semantic concepts and their relationships inherent within the protocols (see Fig. 4B).

**Modeling** We transform the protocols into a vectorized feature space,  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , with each  $\mathbf{x}_i$  encoding operation patterns in a one-hot vector format (refer to Appx. C.3). The objective is to model latent semantic constraint assignments,  $\mathcal{W} = \{w_1, \dots, w_N\}$ , where each  $w_i$  is an assignment of atomic semantic constraint on  $\mathbf{x}_i$ . To note, according to the definition, the size of the semantics set  $\Lambda$  is not fixed and grows with the data.

**Iteration** Utilizing a DPMM facilitates the non-parametric spectral clustering of these feature vectors into groups of unique atomic semantic constraints. Each iteration in the DPMM process involves Gibbs Sampling for estimating the posterior of model’s parameters (see Fig. 4D). This clustering not only identifies distinct semantic operations but also adapts to the complexity and variability of semantic patterns across different domain corpora. Refer to Appx. C.4 for details.

### 3.4 Discussion

We identify the commonalities and distinctions between the DSLs designed by AutoDSL corresponding to the five experimental science domains (see Fig. 4H). We find that the five domains share a majority of syntactic constraints, such as `Memory` and `Branch` (see Fig. 4F), while differing in other constraints, such as `Function-procedure` and `Loop` (see Fig. 4G). This implies that the domain specificity as an experimental protocol dominates that of subject, regarding the syntax.

Different from syntactic constraints, the five domains vary significantly on semantic constraints, including fine-grained concepts such as operations, reagents, and conditions (see Fig. 4E). This implies that the domain specificity of knowledge ontology is dominant regarding the semantics.

## 4 Constraint design assessment

In this section, we evaluate the quality of the DSLs automatically designed by our AutoDSL framework across the five domains. We first discuss the principles for the assessment, and then conduct quantitative and qualitative analysis accordingly.

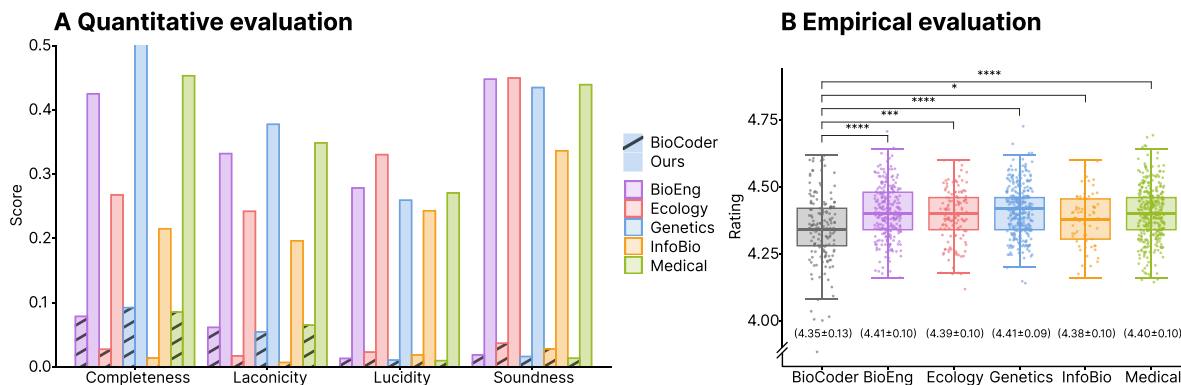


Figure 5: **Constraint design results on 5 experimental science domains.** (A) Quantitative evaluation on the 5 DSL-based constraints and the BioCoder baseline. (B) Empirical evaluation on the 5 DSL-based constraints and the BioCoder baseline.

#### 4.1 Domain corpora

We compile a dataset of 16,194 experimental protocols across five domains: Genetics (8794 protocols), Medical and Clinical Research (7351, “Medical” for abbreviation), Ecology and Environmental Research (812, “Ecology”), Bioengineering (3597, “BioEng”), and Computational Biology (348, “InfoBio”), with minimal overlap between them. This diverse dataset, essential for testing our methodology, reflects the unique syntactic and semantic requirements of each domain. Please refer to [Appx. D.1](#) for details on data collection.

#### 4.2 What makes a good DSL?

We leverage a systematic approach to gain insight into what constitutes the appropriate objectives of DSLs, employing the frameworks introduced by [Guizzard \(2013\)](#) for quantitative and [Karsai et al. \(2009\)](#) for empirical assessment.

**Quantitative evaluation** We check the mapping from ontology elements in the reference model, *i.e.*, concepts and relations in the domain corpus, to DSL constructs of constraints, focusing on four criteria: *soundness*, *lucidity*, *completeness*, and *laconicity*. An ideal DSL should guarantee that (i) all ontology elements are mapped to the construct set for **soundness**; (ii) every ontology element is mapped to only one construct for **lucidity**; (iii) the construct set does not possess any redundancy beyond the ontology elements for **completeness**; (iv) every construct refers to only one ontology element for **laconicity**.

**Empirical evaluation** We take the design guidelines of DSL as questions for assessing the resulting DSLs by AutoDSL, from a user-centric perspective. These questions range from the objective consistency for machine interpretation to the subjective complexity for user understanding. Specifically, the DSL should maintain (i) **simplic-**

**ity**, *i.e.*, being easy to understand both syntactically and semantically; (ii) **clarity**, *i.e.*, pruning unnecessary space for generalization; and (iii) **conciseness**, *i.e.*, avoiding redundant concepts and relations. Please refer to [Appx. E.2](#) for details.

#### 4.3 Quantitative evaluation

**Method** To quantitatively evaluate the effectiveness of DSL-based constraints in protocol representation, we leverage maximum-recall domain-specified ontology knowledge extracted from domain-specific corpora. Such ontology, represented as a graph consisting of concept-relation triplets, serves as the groundtruth for our assessment on the DSLs designed by AutoDSL.

**Results** The analysis on the DSLs designed by AutoDSL, comparing with BioCoder across several domains, is detailed in [Fig. 5A](#). Specifically, our DSLs adeptly model 43.47% of the concepts with precise constraints, *i.e.*, soundness, of which 25.93% showcase a direct one-to-one correspondence, *i.e.*, lucidity. Furthermore, the completeness metric reveals that 50.51% of DSL constraints correspond with specific domain concepts, and 37.74% of constraints are uniquely aligned with a single domain concept, *i.e.*, laconicity. In contrast, BioCoder demonstrates significantly lower performance metrics (lucidity: 1.05%, soundness: 1.61%, laconicity: 5.46%, completeness: 9.22%). This trend of a 5-to-20-fold improvement in the resulting DSLs over BioCoder on each metric is consistent across the five domains, as highlighted in our results.

**Discussion** The quantitative evaluation underscores the superiority of the DSLs designed through AutoDSL over BioCoder, the established DSL hand-crafted by domain experts, in terms of four quantitative metrics. Despite the inherent specificity of the protocols, this assessment high-

lights that a compact subset of constraints can be precisely defined and extracted. These results validate our framework’s ability to model constraints that more accurately and effectively encapsulate domain-specified procedural knowledge.

#### 4.4 Empirical evaluation

**Method** We institute an automatic evaluation framework leveraging a state-of-the-art LLM, *e.g.*, GPT-4, augmented with a novel Question-answering System (QA System) designed to simulate the analytical capabilities of human domain experts. This QA System is engineered using an index vector embedding technique to integrate domain-specific corpora from the five domains into the LLM. To facilitate nuanced and domain-aware responses, we employed advanced prompt-engineering strategies, including the Chain-of-Thought (CoT) technique (Wei et al., 2022), enabling the QA System to generate assessments that closely mirror those of a domain expert. The system’s empirical evaluation was predicated on its performance across a set of 50 meticulously designed questions, as delineated in Appx. E.2.

**Meta-evaluation** Prior to deploying this automated evaluation mechanism across our DSLs, we conduct a preliminary meta-evaluation to ascertain the congruence between the QA System’s assessments and those of human domain experts, following the routine of automatic evaluation (Schuff et al., 2023). This process involves a comparative analysis on four subsets extracted from our DSLs and BioCoder, with both human domain experts ( $N = 3$ ) and the QA System providing ratings. We observe no significant evidence supporting that human experts and the QA System rate differently ( $t(99) = -1.282, \mu_d \neq 0, p = .202$ ), validating the application of our automated evaluation framework in assessing the resulting DSLs.

**Results** In the ensuing phase, the QA System appraises the five DSLs, yielding ratings and confidence scores. A subsequent analysis employing paired samples T-tests (see Fig. 5B;  $t(49) = 3.487, \mu_d \leq 0, p < .005$ ) substantiated that our resulting DSLs exhibit quality metrics on par with those of BioCoder — the benchmark DSL meticulously crafted using C/C++.

**Discussion** The empirical evaluation not only underscores the viability of our AutoDSL framework in automating the design of domain-specific languages, but also highlights its potential to

match or even surpass the quality of manually engineered solutions like BioCoder. It is primarily attributed to the design decisions to optimize a broad language feature set for compact syntactic constraints; and to tailor operations for the domain for precise semantic constraints. These strategies establish a clean and compact DSL-based constraint that adheres to domain conventions.

## 5 Constraint utility assessment

In this section, we evaluate the utility of leveraging DSL-based constraints for the representation of new coming experimental protocols across the five domains, considering both syntax and semantics. Please refer to Tab. 1 for a demonstration.

### 5.1 Materials

Our dataset includes 186 new coming protocols, meticulously collected from recent experiments by domain experts across the specified domains. These protocols, which have been validated for accuracy, serve as a testbed for evaluating the performance of DSL-based constraints on **unseen** data during the design phase. We identify the challenges with four syntactic features — Imperative control flows, Type system, Concurrent, and Reactive model. There are also three semantic errors — Action undefinition, Parameter omission, and Parameter under-specification.

For syntactic constraints, we determine the success rate by calculating the proportion of samples where the target challenging syntactic features are accurately constrained. Similarly, for semantic constraints, the success rate is assessed by the proportion of samples where the DSL successfully identifies and resolves semantic errors. For fair comparison, we transform the varied outputs from different approaches to a unified JSON-style representation. For strings referring to names of instances, we relax the exact-match criteria to similarity-based score (Papineni et al., 2002).

### 5.2 Methods

We employ a multi-dimensional approach to assess the utility of DSL-based constraints, comparing them with several alternative methods. This maintains the integrity of an end-to-end workflow with minimal adjustments to the LLM.

**Our DSL-based constraint & LLM** Incorporating DSL as an external constraint involves using it as an interpreter for the programs gener-

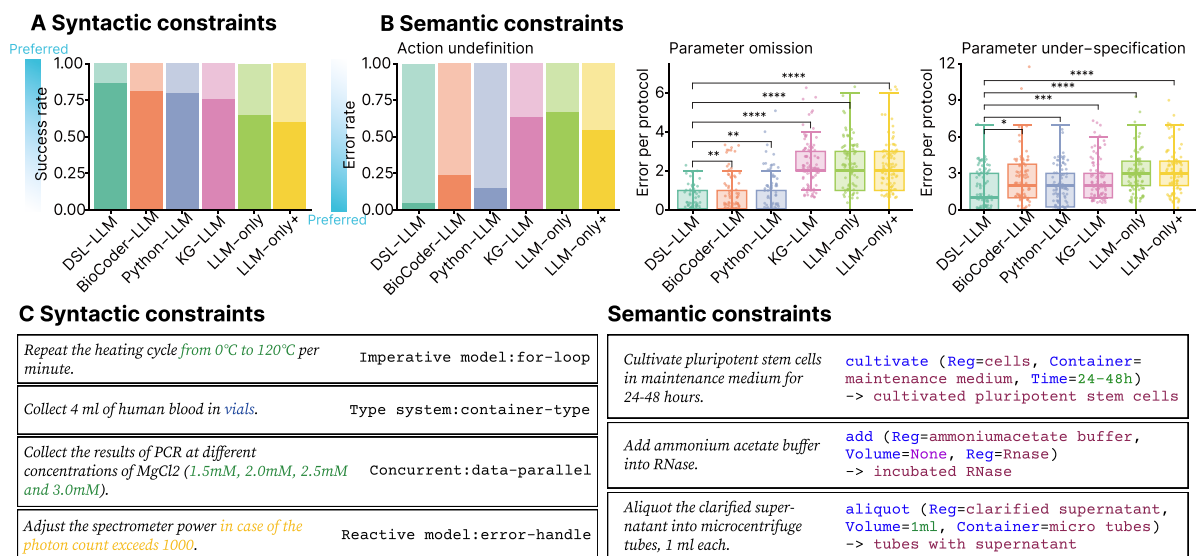


Figure 6: **Constraint utility assessment results on the five experimental science domains between DSL-based constraint and alternative models.** (A) Results on syntactic constraint utility assessment (**higher is preferred**). (B) Results on semantic constraint utility assessment (**lower is preferred**). (C) New coming protocols represented with syntactic or semantic constraints.

ated by the LLM. Following protocols suggested by recent research (Gao et al., 2023; Zhang et al., 2023), we prompt the LLM with DSL syntax grammar and semantics operation set with running examples, directing it to translate procedural texts into corresponding programs. These programs are then verified and potentially corrected using the DSL’s syntactic and semantic constraints. This approach minimizes external dependencies, striving for a seamless plug-and-play integration between DSLs and LLMs. We utilize GPT-3.5 as the backbone model, resulting in DSL-LLM.

**Programming language & LLM** As comparative baselines, we explore the use of another DSL, BioCoder, and a general-purpose programming language, Python, with GPT-3.5 as the base model. These methodologies are denoted as BioCoder-LLM and Python-LLM, respectively. The main variance lies in the adoption of one-shot generation for fair comparison, given GPT’s pre-existing familiarity with both BioCoder and Python’s coding paradigms.

**Structural representation & LLM** We also examine the effectiveness of elementary structural knowledge representation, specifically Knowledge Graph (KG), as a simpler alternative. This method, KG-LLM, leverages entity-relation extraction techniques for knowledge structuring.

**LLM only** Finally, we assess the capability of naive LLMs operating without any structural constraints, particularly using GPT-3.5 and GPT-4, to gauge the impact of LLM advancements alone, named after LLM-only and LLM-only+.

### 5.3 Results

**Syntactic constraints** Our investigation into syntactic constraints shows that DSL-LLM surpasses all alternative approaches in terms of success rate across the five domains (see Fig. 6A). In the analysis of syntactic constraints across five domains, DSL-LLM achieves a success rate of 93.5%, which significantly outperforms the counterparts without programming language representation ( $\chi^2(1) = 3.979, p < .05$ ), and also outperforms the counterparts with general syntactic constraints. This substantiates the efficacy of these domain-specific syntactic constraints in meticulously constraining non-trivial protocols.

**Semantic constraints** The assessment of semantic constraints reveals that DSL-LLM outperforms both the strong and weak baselines (see Fig. 6B) regarding the three types of errors. In the context of Genetics domain, DSL-LLM yields a success rate of 93.7% in addressing semantic errors, significantly outperforms alternative methods, as indicated by statistical evidence ( $\chi^2(1) = 8.378, p < .005$  in action undefinition;  $t(185) = -3.215, \mu_d < 0, p < .005$  in parameter omission; and  $t(185) = -2.164, \mu_d < 0, p < .05$  in parameter under-specification). These underscore the domain-specific semantic constraints’ capability in enhancing accuracy of protocol representation, which is crucial for successful experiments.

### 5.4 Discussion

The performance of alternative approaches basically aligns with our expectation (see Fig. 6C).



Table 1: **Showcases of protocol representations with all approaches.** (OP: original protocol; GT: ground truth; L+: LLM-only+; each abbreviation consists of the two initials from both sides of the dash in each approach name)

| OP | <b>RNA Determination:</b> <i>Add ammonium acetate buffer and RNaseT2, then incubate.</i>                      | <b>Bile Processing:</b> <i>Use a centrifuge tube to spin at 3000g for 10 min at 4°C.</i>                      | <b>Bile Collection:</b> <i>Collect bile into a sterile collection bottle.</i>                 |
|----|---|---|---|
| GT | ADD: [[Reg: ammonium acetate buffer], [Container: None], [Volume: None], [Reg: RNaseT2]] -> incubated RNaseT2 | SPIN: [[Force: 3000g], [Time: 10min], [Temperature: 4°C], [Container: centrifuge tube]] -> centrifuged sample | COLLECT: [[Reg: bile], [Volume: None], [Container: sterile collection bottle]] -> bile sample |
| DL | ADD: [[Reg: ammonium acetate buffer], [Container: None], [Volume: None], [Reg: RNaseT2]] -> incubated RNaseT2 | SPIN: [[Force: 3000g], [Time: 10min], [Container: centrifuge tube], [Temperature: 4°C]] -> centrifuged sample | COLLECT: [[Reg: bile], [Volume: None], [Container: sterile collection bottle]] -> bile sample |
| BL | ADD: [[Reg: ammonium acetate buffer], [Volume: None], [Reg: RNaseT2]] -> incubated RNaseT2                    | SPIN: [[Container: tube], [Force: 3000g], [Time: 10min], [Temperature: 4°C]] ->                               | COLLECT: [[Container: sterile collection bottle]] -> centrifuged pellet                       |
| PL | ADD: [[Reg: ammonium acetate buffer], [Volume: None], [Reg: RNaseT2]] -> incubated RNaseT2                    | SPIN: [[Force: 3000g], [Time: 10min], [Temperature: 4°C]] -> centrifuged sample                               | CENTRIFUGE: [[Container: bottle]] -> centrifuged bile   |
| KL | : [] ->   | STEP1: [[Force: 3000g], [Time: 10min], [Temperature: 4°C]] ->   | : [] ->   |
| LO | : [] ->   | STEP: [[Device: centrifuge], [Force: 3000g], [Time: 10min], [Temperature: 4°C]] ->                            | CENTRIFUGE: [[Container: sterile collection bottle]] -> centrifuged bile                      |
| L+ | ADD: [[Reg: ammonium acetate buffer], [Reg: RNaseT2]] -> incubated RNaseT2                                    | SPIN: [[Device: tube], [Time: 10min], [Force: 3000g], [Temperature: 4°C]] -> sample                           | CENTRIFUGE: [[Reg: bile], [Container: sterile collection bottle]] -> collected bile           |

Among them, DSL-LLM demonstrates the highest performance in addressing both syntactic and semantic constraints. While BioCoder-LLM and Python-LLM, the strong baselines, are outperformed by DSL-LLM in most times, they still exhibit substantial advantages over other baselines. This phenomenon can be attributed to the merits shared by programming languages, such as the ability to represent structural knowledge at various levels of granularity. Although not as effective as semantics constraints, which exactly define the structures and legal ranges for the space of operations, reagents, and conditions, this type of representation still constrains the potential search space to some extent. On the other hand, the flattened structural representation of KGs cannot provide the same level of expression capacity for KG-LLM as programming languages. Despite the expectation that GPT-4 would be much more capable than GPT-3.5, the performances of LLM-only and LLM-only+ are comparable, suggesting that a pure text representation may not be suitable for processing complicated procedures like protocols. The relative success of DSL-LLM indicates the potential of DSLs as external constraints for LLMs.

## 6 General discussions

In this work, we present the AutoDSL framework as a proof-of-concept to facilitate the automation of designing DSL-based constraints across various domains. Through both quantitative and qualitative evaluations of the DSLs designed by AutoDSL in five distinct domains, we demonstrate

its capability as an auxiliary module for LLM.

**Rationale behind DSL** The decision to leverage DSL for constraint representation is rooted in several considerations. Primarily, DSLs adeptly capture domain-specified syntactic and semantic constraints, aligning well with the hardness of representing protocols with complicated control flows and operations. Moreover, DSLs leverage a deterministic verification mechanism derived from general-purpose programming languages, offering a robust means of imposing constraints on the inherently nondeterministic outputs of LLMs. Additionally, DSLs are user-friendly to both humans and machines, maintaining a minimal set of language features that facilitate ease of adoption.

**No universal constraint** It is unrealistic to expect a one-size-fits-all protocol constraint applicable across varied domains. For a constraint system to accurately delineate the execution space of a particular domain, it must encompass domain-specific syntax and semantics. Though it is possible to devise a comprehensive set of constraints that covers the requirements of all conceivable domains, such an approach would yield a constraint system of prohibitive complexity, rendering it impractical for end-users. Conversely, simplifying this universal constraint to enhance user-friendliness inevitably compromises its capability, leading to the expressivity-complexity dilemma (Abelson and Sussman, 1996). Rather than seeking an elusive generality, focusing on domain-specific constraint development and striving for the automation of this process may offer a pragmatic way to circumvent this dilemma.

## Ethics statement

**Human participants** The meta-evaluation and data annotation included in this work has been approved by the Institutional Review Board (IRB) of Peking University. We have been committed to upholding the highest ethical standards in conducting this study and ensuring the protection of the rights and welfare of all participants. Every domain expert was paid on a wage of \$22.5 per hour for participating in the meta-evaluation and data annotation. We have obtained informed consent from all participants, including clear and comprehensive information about the purpose of the study, the procedures involved, the risks and benefits, and the right to withdraw at any time without penalty. Participants were also assured of the confidentiality of their information. Any personal data collected, including name, age, gender, institution, and education background, was handled in accordance with applicable laws and regulations.

**Corpora collection** We carefully ensure that all experimental protocols incorporated into our corpora strictly adhere to open access policies, governed by the Creative Commons license. This approach guarantees full compliance with copyright and intellectual property laws, eliminating any potential infringement or unauthorized use of protected materials. By exclusively utilizing resources that are freely available and legally distributable, we uphold the highest standards of ethical conduct in research, fostering an environment of transparency and respect for the intellectual property rights of others. This commitment ensures that our work not only advances the frontiers of knowledge but does so in a manner that is both legally sound and ethically responsible.

## Limitations

As a proof-of-concept work, the design and evaluation of AutoDSL come with limitations, leading to further investigations:

- We majorly consider the imperative programming model as the prior model for DSL design in the work. This raises the question of whether incorporating alternative programming paradigms, such as functional and object-oriented models, could enhance the representation of complex entities within protocols, particularly the properties of reagents.
- AutoDSL's current framework outputs only syntactic and semantic constraints, lacking an ex-

PLICIT production system. This leads to the question of whether it is feasible to augment the AutoDSL framework to autonomously generate both a production system and a constraint system, leveraging domain-specific corpora and pre-existing knowledge on programming languages. Such an enhancement could significantly improve DSL's potential on planning.

- To ensure a fair comparison and to underscore the plug-in capability of the developed DSLs, only minimal prompt engineering is applied in protocol processing. This posits the potential for developing tools that could more effectively intertwine DSLs with LLMs.
- Notably, the DSL-InfoBio's quantitative evaluation outcomes are markedly inferior to those of its DSL counterparts. Considering the notably smaller size of the InfoBio domain corpus relative to other domain-specific corpora, this prompts an investigation into the potential correlation between the scale of a domain corpus and the quality of the resulting DSL.

With many questions unanswered, we hope to explore more on automated design of DSL-based constraints for procedural representation.

## Reproducibility

The project page with supplementary files for reproducing the results of this paper is available at <https://autodsl.org/procedure/papers/acl24shi.html>.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 91948302. Part of the authors are visiting students at Peking University during this work. In particular, Z. Bi is visiting from Huazhong University of Science and Technology, X. Wei is visiting from Beijing Jiaotong University. The authors would like to appreciate these institutes, and also Prof. Wenjuan Han for her support.

## References

- Harold Abelson and Gerald Jay Sussman. 1996. *Structure and interpretation of computer programs*. The MIT Press.
- Vaishnavi Ananthanarayanan and William Thies. 2010. Biocoder: A programming language for standardizing and automating biology protocols. *Journal of Biological Engineering*, 4(1):1–13.

- Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604).
- Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.
- Bruce G Buchanan and Edward A Feigenbaum. 1978. Dendral and meta-dendral: Their applications dimension. *Artificial Intelligence*, 11(1-2):5–24.
- Noam Chomsky. 1957. *Syntactic Structures*. Mouton de Gruyter.
- Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. 2021. Aligning actions across recipe graphs. In *Annual Conference on Empirical Methods in Natural Language Processing*.
- Edward A Feigenbaum. 1981. Expert systems in the 1980s. *State of the art report on machine intelligence. Maidenhead: Pergamon-Infotech*, 23.
- Martin Fowler. 2010. *Domain-specific languages*. Pearson Education.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*.
- Giancarlo Guizzardi. 2013. Ontology-based evaluation and design of visual conceptual modeling languages. *Domain Engineering: Product Lines, Languages, and Conceptual Models*, pages 317–347.
- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 1996. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Longman Publishing Co., Inc.
- Jiani Huang, Ziyang Li, Binghong Chen, Karan Samel, Mayur Naik, Le Song, and Xujie Si. 2021. Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In *Advances in Neural Information Processing Systems*.
- Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*.
- Gabor Karsai, Holger Krahn, Claas Pinkernell, Bernhard Rumpe, Martin Schindler, and Steven Völkel. 2009. Design guidelines for domain specific languages. In *OOPSLA Workshop on Domain-Specific Modeling (DSM' 09)*.
- Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Annual Conference on Empirical Methods in Natural Language Processing*.
- Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.
- Marjan Mernik, Jan Heering, and Anthony M Sloane. 2005. When and how to develop domain-specific languages. *ACM Computing Surveys (CSUR)*, 37(4):316–344.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.
- Yoshio Momouchi. 1980. Control structures for actions in procedural texts and pt-chart. In *COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics*.
- Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):1–9.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.
- Paolo Pareti, Ewan Klein, and Adam Barker. 2014a. A semantic web of know-how: linked data for community-centric tasks. In *Proceedings of the International Conference on World Wide Web*.
- Paolo Pareti, Benoit Testu, Ryutaro Ichise, Ewan Klein, and Adam Barker. 2014b. Integrating know-how into the linked data cloud. In *Knowledge Engineering and Knowledge Management: 19th International Conference*.
- Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *Conference on Computer Vision and Pattern Recognition*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits

- of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Bertrand Russell. 1923. Vagueness. *The Australasian Journal of Psychology and Philosophy*, 1(2):84–92.
- Stuart J Russell and Peter Norvig. 2010. *Artificial intelligence a modern approach*. Prentice Hall Press.
- Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural Language Processing using very large Corpora*.
- Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, pages 1–24.
- Yu-Zhe Shi, Shiqian Li, Xinyi Niu, Qiao Xu, Jiawen Liu, Yifan Xu, Shiyu Gu, Bingru He, Xinyang Li, Xinyu Zhao, et al. 2023. PersLEARN: Research training through the lens of perspective cultivation. In *Annual Meeting of the Association for Computational Linguistics*.
- Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Annual Conference on Empirical Methods in Natural Language Processing*.
- Alfred Tarski. 1946. *Introduction to Logic and to the Methodology of Deductive Sciences*. Dover Publications.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.
- Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*.
- Ziqi Zhang, Philip Webster, Victoria S Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically extracting procedural knowledge from instructional texts using natural language processing. In *LREC*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.
- Shuyan Zhou, Pengcheng Yin, and Graham Neubig. 2022a. Hierarchical control of situated agents through natural language. In *Proceedings of the Workshop on Structured and Unstructured Knowledge Integration*.
- Shuyan Zhou, Li Zhang, Yue Yang, Qing Lyu, Pengcheng Yin, Chris Callison-Burch, and Graham Neubig. 2022b. Show me more details: Discovering hierarchies of procedures from semi-structured web data. In *Annual Meeting of the Association for Computational Linguistics*.



## A Additional remarks

### A.1 AutoDSL and expert systems

The expert system is a highly structured representation of knowledge, and has been intensively studied while widely applied in the past few decades. Compared with the statistical model, however, it often requires more hand-crafted effort for development, and the workload increases dramatically with the increasing generality level. This echoes the primary goal of our proposed AutoDSL framework, which makes efforts to formulate the structure of knowledge representation, similar to classic expert systems such as [Buchanan and Feigenbaum \(1978\)](#); [Feigenbaum \(1981\)](#), while adopting the power of data and LLM for domain generality and adaptation.

### A.2 AutoDSL framework and LLM

The current AutoDSL framework is essentially an LLM-in-the-loop approach. In the pre-processing of raw protocols, we take full advantage of LLM for conducting state-of-the-art Information Extraction (IE), obtaining key information from raw text, such as entities of operations, entities of conditions, and relations between different operations, *etc.* Such key information is taken as the input of our core algorithms for syntactic constraint optimization and semantic constraint abstraction.

We would like to reiterate that AutoDSL is **not an alternative** for LLM. The proposed AutoDSL framework is in essence yet another workflow for utilizing LLM to handle protocols with constraints. Compared with pure end-to-end workflow of LLM for producing such constraints, our framework is designed to be an explicit two-stage workflow. In the first stage, the framework automatically designs DSLs for target domains in a bidirectional optimization fashion. In the second stage, the resulting DSLs serve as plug-and-play constraint modules for all kinds of LLMs to handle protocols precisely, preventing them from generating free-formed and non-deterministic procedural interpretations.

## B On the generality of AutoDSL

### B.1 Generalize in the context of experimental protocols

Regarding domains related to experimental sciences, there are syntactic constraints shared by domains and those unique to domains respectively, as we have discussed in [Sec. 3.4](#). These shared syntactic constraints mainly reflect the procedural nature across domains, and unique ones attributed to the composition types and the interruption in procedure execution. Through the following demonstrations, we aim to show that while semantic constraints are more diverse than syntactic constraints, there are still substantial distinctions between syntactic constraints across different domains.

**Shared syntactic constraints across domain** There are some examples of syntactic constraints that are shared across different areas:

- `allocate-statement`: *Inoculate 5 ml LB medium containing 100 µg/ml ampicillin with bacteria.*
- `if-branch`: *If cells are to be treated with PTX, divide them into two groups: one receives 100 ng/ml PTX.*
- `parallel-for`: *Iterate different MgCl<sub>2</sub> concentrations (1.5mM, 2.0mM, 2.5mM, and 3.0mM) to find the optimal concentration for DNA amplification.*
- `temporal-type-declaration`: *Gently shake the reagent for 5 seconds to mix.*

**Unique syntactic constraints of one particular domain** There are some examples of syntactic constraints that are unique to particular domains respectively:

- In BioEng, `container-type-declaration`: *Resuspend the pellet in PBS to refill the **tube**.*
- In Ecology, `string-type-declaration`: *Close the soundproof chamber as instruction manual (see "[https://—](#)").*

- In InfoBio, raise-statement: *If unbalanced spectral peaks, noisy data, and frame loss occur during recording, ensure to subtract background noise and adjust for hemodynamic changes.*

**Unique semantic constraints of one particular domain** There are some examples of semantic constraints that are unique to particular domains respectively:

- In Genetics,
  - DILUTE: Dilute the Spike-inAmpR fragment to a concentration of 65.6 nM.
  - ATTACH: Attach a Slide-A-Lyzer Float Buoy to the top (single) dialysis clip.
  - DELETE: Detect the flow-through samples from each step with SDS-PAGE gel.
- In Medical,
  - ASPIRATE: Aspirate the DPBS and add 40  $\mu$ 03bcL of Sigma Lysis Solution for Blood.
  - ANESTHETIZE: Anesthetize the rat intraperitoneally (i.p.).
  - DISSECT: Dissect the fly brain under a stereomicroscope with light sources.
- In Ecology,
  - STERILIZE: Sterilize the seeds with 5% (v/v) sodium hypochlorite.
  - QUANTIFY: Quantify the eggs by placing washed eggs in a 10-mL graduated cylinder.
  - CALIBRATE: Calibrate the motion sensor using a two-axis actuator.

## B.2 Generalize beyond experimental protocols

AutoDSL focuses on a specialized form of natural language: procedural text, and is designed to formalize this procedural text, offering a clear execution trace and a readily verifiable interface. We select experimental sciences experiments for their demands on high interpretability, stringent execution, and adaptable planning. This methodology can be easily applied to new domains beyond experimental sciences, where such rigorous conditions are not necessary, including culinary recipes. Examples of various syntactic constraints are provided in Tab. A1.

Table A1: Showcases of syntactic constraints in culinary recipes

| Constraint                     | Original recipe text                  |
|--------------------------------|---------------------------------------|
| integer-type-declaration       | 8 [ounces] white fresh {pasta}        |
| device-type-declaration        | Yield: 2 plates                       |
| floatingpoint-type-declaration | 1/3 [cup] red {wine}                  |
| temporal-type-declaration      | After  2 minutes  more, add the beef. |

We can also effectively design a DSL with such a corpus where actions happen in a kitchen instead of a lab. We showcase a recipe example adopted from the Corel<sup>1</sup> DSL as follows.

```

1 Pasta Bolognese
2
3 Yield: 2 plates
4
5 Ingredients:
6
7 - 8 [ounces] white fresh {pasta}
8 - 1 [floz] olive {oil}
9 - 1/4 [ounce] {garlic}; minced
10 - 4 [ounces] {onions}; chopped
11 - 4 [ounces] shallow fried {beef}; minced
12 - 1 - 1 1/2 [ounce] lean prepared {bacon}
13 - 1/3 [cup] red {wine}

```

<sup>1</sup>Visit <https://fse.studenttheses.ub.rug.nl/25731/> for documentation.

```

14 - 150 [gram] raw {carrots}; thinly sliced
15 - 2/3 [ounce] concentrated {tomato puree}
16 - 4 [ounces] red {sweet pepper}; cut julienne
17 - 1 [ounce] {parmesan} cheese
18
19 Instructions:
20
21 Add the @oil@ to a large saucepan, heat to <300 F>, and saute the @onions@.
22
23 After |2 minutes|, add the @garlic@.
24
25 Keep on medium to high heat, and don't stir.
26
27 After |2 minutes| more, add the @beef@.
28
29 Fry the @bacon@ in a separate pan, on high heat.
30
31 Remove liquified fat when done.
32
33 Boil @pasta@ in a medium pan, until al dente (~|8 minutes|).
34
35 Drain when done.
36
37 Once the @beef@ is done, add the @carrots@, @sweet pepper@ and @tomato puree@.
38
39 Slowly add the @wine@ as well, to not lower the temperature.
40
41 Let it simmer (but not boil) for |5-10 minutes|.
42
43 Add the @bacon@ to the large saucepan.
44
45 Serve with grated @parmesan@ cheese.

```

Based on the example, it is trivial to figure out that the constraints on cooking recipes are essentially in a subset of those scientific procedures. The cooking procedures mainly focus on sequentially executing the actions without switching their execution orders. Also, most of the culinary ingredients are processed in single-thread fashion without duplication for different experimental groups. Hence, the Corel DSL does not require some syntactic constraints for non-linear control flows, such as `loop`, `branch`, and `split`, while these constraints are dominant in DSLs for protocols. AutoDSL will adaptively exclude these unnecessary features when optimizing the DSL for culinary procedures given the corpus of recipes. Regarding semantics constraints, there are fewer shared semantics on operations, ingredients, and conditions between cooking and experimental sciences compared to those semantics shared between different domains within experimental sciences. However, this increment of semantic diversity would not yield significant challenge for AutoDSL, as our semantic reduction does not rely on domain transfer.

### B.3 Extended discussions

When considering both syntactic and semantic constraints across different domains, we observe that there are significant differences within the context of experimental sciences. However, when looking at a more general context that encompasses all possible procedural knowledge worldwide, these differences tend to converge. This suggests that DSLs may have a **hierarchical** structure in the general context. The shared features of different DSLs are more common closer to the root, such as the operation `incubate` shared by Genetics and Ecology in experimental sciences, and the operation `add` shared by Genetics, Ecology, and cooking. These common DSL constructs possess general semantics. On the other hand, as we move closer to the leaf, the majority of DSL constructs become specialized for unique domains. This

parallels the spectrum of *naturalism* in sciences (Shi et al., 2023), ranging from a general community with convergent high-level concepts to a specific community with divergent low-level knowledge.

## C Implementation details

### C.1 Implementation of the syntax prior model

We employ the CFG generally for designing modern imperative programming language as the prior model  $p(\mathcal{Z}|\mathcal{S})$  for syntax optimization.

```

1 <program> ::= <statements>
2
3
4 <statements> ::= <statement>
5                 | <statement> <statements>
6
7 /* All support */
8 <statements> ::= <imperative-model>
9                 | <runtime-error-handling>
10                | <type-system>
11                | <concurrent>
12                | <react>
13
14
15 /* Imperative Model */
16 <imperative-model> ::= <if-statement>
17                     | <loop-statement>
18                     | <jump-statement>
19                     | <memory-management>
20                     | <function-procedure>
21                     | <arithmetic-expression>
22                     | <logical-expression>
23                     | <assignment-expression>
24
25 <if-statement> ::= "if" "(" <expression> ")" "{" <statements> "}"
26                | "if" "(" <expression> ")" "{" <statements> "}" "else" "{" <
27                statements> "}"
28
29 <loop-statement> ::= "While" "(" <expression> ")" "{" <statements> "}"
30                | "For" "(" <assignment-expression> ";" <expression> ";" <
31                assignment-expression> ")" "{" <statements> "}"
32
33 <jump-statement> ::= "break"
34                | "continue"
35
36
37 <function-procedure> ::= "Call" <identifier> "(" <arguments> ")"
38                | "Function" <identifier> "(" <parameters> ")" "{" <
39                statements> "}"
40
41
42 <memory-management> ::= "allocate" <type> <identifier>
43                | "deallocate" <identifier>
44
45 <assignment-expression> ::= <identifier> "=" <expression>
46
47 <arithmetic-expression> ::= <expression> <arithmetic-operator> <expression>
48
49 <logical-expression> ::= <expression> <logical-operator> <expression>

```



```

45
46 /* Runtime Error Handling */
47 <runtime-error-handling> ::= <raise-stmt>
48                               | <resolve-stmt>
49
50 <raise-stmt> ::= "raise" "(" <expression> ")"
51
52 <resolve-stmt> ::= "try" "{" <statements> }" "catch" "(" <identifier> ")" "{" <
53   statements> }"
54
55 /* Type System */
56 <type-system> ::= <data-type>
57                 | <class-type>
58                 | <domain-specified-type>
59
60 <domain-specified-type> ::= "time"
61                             | "reagent"
62                             | "device"
63                             | "container"
64                             | <scientific-type>
65
66 <data-type> ::= "int"
67               | "float"
68               | "bool"
69               | "string"
70               | "set" "<" <type> ">"
71               | "dict" "<" <type> "," <type> ">"
72               | "vector" "<" <type> ">"
73
74 <class-type> ::= "class" <identifier> "{" <class-body> }"
75
76 <kind-type> ::= "type" <identifier> "=" <type>
77
78 <class-body> ::= <class-members>
79
80 <class-members> ::= <class-member> | <class-member> <class-members>
81
82 <class-member> ::= <variable-declaration>
83                 | <method-declaration>
84
85 <variable-declaration> ::= <assignment-expression>
86
87 <method-declaration> ::= <function-procedure>
88
89 /* Concurrent Programming */
90 <concurrent> ::= <data-parallel>
91               | <message-passing>
92
93 <data-parallel> ::= "parallelFor" "(" <parallel-range> ")" "{" <statements> }"
94                 | "parallelMap" "(" <collection> "," <function> ")"
95
96 <parallel-range> ::= "range" "(" <expression> "," <expression> ")" // Define start,
97                       end of range

```

```

97 <collection> ::= <identifier> // Reference to a collection of data, e.g., array,
    list
98
99 <function> ::= <identifier> // Reference to a function to apply in parallel
100
101 <message-passing> ::= "spawnProcess" "(" <process-function> ")"
102                     | "sendMessage" "(" <process-identifier> "," <message> ")"
103                     | "receiveMessage" "(" <message-type> ")"
104
105 <process-function> ::= <identifier>
106
107 <process-identifier> ::= <identifier>
108
109 <message> ::= <expression>
110
111 <message-type> ::= <type>
112
113
114 /* React Model */
115 <react> ::= <event-stmt>
116           | <response-stmt>
117
118 <event-stmt> ::= "emit" "(" <event> ")"
119
120 <response-stmt> ::= "on" "(" <event> ")" "{" <statements> "}"
121
122 /* Auxiliary Definitions */
123 <type> ::= <data-type> | <class-type> | <kind-type> | "void"
124
125 <parameters> ::= <empty> | <parameter> | <parameter> "," <parameters>
126
127 <parameter> ::= <identifier> ":" <type>
128
129 <arguments> ::= <empty> | <expression> | <expression> "," <arguments>
130
131 <arithmetic-operator> ::= "+" | "-" | "*" | "/"
132
133 <logical-operator> ::= "&&" | "||" | "!"
134
135 <identifier> ::= <letter> (<letter> | <digit>)*
136
137 <letter> ::= "A" | "B" | ... | "Z" | "a" | "b" | ... | "z"
138
139 <digit> ::= "0" | "1" | "2" | ... | "9"
140
141 <event> ::= <identifier>

```

## C.2 Implementation of the prior belief function

$p(\theta_k | \mathbf{c}_i)$  serves as a prior and represents the probability that programming language features  $\theta_k$  are present within the experimental protocol  $\mathbf{c}_i$ .

$$p(\theta_k | \mathbf{c}_i) = \begin{cases} \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{INT}_i) & \text{if } \theta_k = \text{interger-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{FLOAT}_i) & \text{if } \theta_k = \text{floatingpoint-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max(\text{Exists}(\text{TRUE}_i), \text{Exists}(\text{FALSE}_i)) & \text{if } \theta_k = \text{boolean-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{STRING}_i) & \text{if } \theta_k = \text{string-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{S}(\text{NOUN}_i, \text{"vector"})\} & \text{if } \theta_k = \text{vector-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{S}(\text{NOUN}_i, \text{"dictionary"})\} & \text{if } \theta_k = \text{dict-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{S}(\text{NOUN}_i, \text{"set"})\} & \text{if } \theta_k = \text{set-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{TIME}_i) & \text{if } \theta_k = \text{temporal-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{BelongTo}(\text{NOUN}_i, \text{Chem})\} = \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{CHEM}_i) & \text{if } \theta_k = \text{reg-type-declaration} \\ 1 & \text{if } \theta_k = \text{device-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{BelongTo}(\text{NOUN}_i, \text{Container})\} = \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{CONTAINER}_i) & \text{if } \theta_k = \text{container-type-declaration} \\ 1 & \text{if } \theta_k = \text{scientific-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"repeat"}) & \text{if } \theta_k = \text{for-loop} \\ \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"repeat"}) \times \text{!Exists}(\text{INT}_i) & \text{if } \theta_k = \text{while-loop} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{IF}_i) & \text{if } \theta_k = \text{if-branch} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{IF}_i) \times \text{Exists}(\text{ELSE}_i) & \text{if } \theta_k = \text{if-else-branch} \\ \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"call"}) & \text{if } \theta_k = \text{function-procedure-call} \\ \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"call"}) & \text{if } \theta_k = \text{function-procedure-declaration} \\ Q(\text{while-loop} | \mathbf{c}) \times Q(\text{if-branch} | \mathbf{c}_i) & \text{if } \theta_k = \text{break-statement} \\ 0 & \text{if } \theta_k = \text{continue-statement} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{BelongTo}(\text{NOUN}_i, \text{Chem}), \text{BelongTo}(\text{NOUN}_i, \text{Container})\} & \text{if } \theta_k = \text{allocate-statement} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{ADD}_i) & \text{if } \theta_k = \text{add-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{MINUS}_i) & \text{if } \theta_k = \text{minus-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{MULTIPLY}_i) & \text{if } \theta_k = \text{multi-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{DEVIDE}_i) & \text{if } \theta_k = \text{devid-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{AND}_i) & \text{if } \theta_k = \text{and-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{OR}_i) & \text{if } \theta_k = \text{or-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{NOT}_i) & \text{if } \theta_k = \text{not-arithmetic-operator} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{EQUAL}_i) & \text{if } \theta_k = \text{assignment-expression} \\ \max_{i=1}^{L(\text{opc})} \max\{\text{S}(\text{NOUN}_i, \text{"error"})\} & \text{if } \theta_k = \text{raise-statement} \\ \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"resolve"}) \times \max\{\text{S}(\text{NOUN}_i, \text{"error"})\} & \text{if } \theta_k = \text{resolve-statement} \\ \exists x \in \text{Chem} : \sum_{i=1}^{L(\text{opc})} \text{Exists}(x \in \text{NOUN}_i) \geq \epsilon & \text{if } \theta_k = \text{class-type-declaration} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{PERSONS}_i) & \text{if } \theta_k = \text{spawn-process} \\ Q(\text{spawn-process} | \mathbf{c}) \times \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"say"}) & \text{if } \theta_k = \text{send-message} \\ Q(\text{spawn-process} | \mathbf{c}) \times \max_{i=1}^{L(\text{opc})} \text{S}(\text{opc}_i, \text{"say"}) & \text{if } \theta_k = \text{receive-message} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{WHEN}_i) & \text{if } \theta_k = \text{event-statement} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(\text{WHEN}_i) & \text{if } \theta_k = \text{response-statement} \\ \max_{i=1}^{L(\text{opc})} \text{Exists}(L(\text{NUM}_i) \geq \epsilon) & \text{if } \theta_k = \text{parallel-for} \\ \max_{i=1}^{L(\text{opc})-1} \max_{j=i+1}^{L(\text{opc})} \text{Exists}(\text{CHEM}_i) \times \text{Exists}(\text{CHEM}_i = \text{CHEM}_j) \times \\ \text{!Exists}(\text{PRON}_i) \times \text{!Exists}(\text{PRON}_j) \times (1 - \text{S}(\text{Sentence}_i, \text{Sentence}_j)) & \text{if } \theta_k = \text{parallel-map} \end{cases}$$

In the context of our prior belief function, `Chem` and `Container` are predefined sets. The set `Sentence` includes all possible sentences, each of which begins with an opcode. These opcodes together constitute the set `"opc."` In our algorithm, we assign  $\epsilon$  with 4 and assign  $\delta$  with 0.5.

We derive these various sets from filtering based on specific rules.

For pre-processing, We utilize regular expressions to match a variety of elements, such as integers, floating-point numbers, boolean values, strings, and control flow elements. Then, we employ speech tagging to label nouns and pronouns (Schmid, 1999). Afterwards, we conduct entity recognition for the identification of chemical elements.

To determine word similarity, we employ word embedding model to calculate cosine similarities (Mikolov et al., 2013). To assess sentence similarity, pairwise word similarity scores are averaged

across the words in the sentences. Furthermore, we use GPT-3.5 to extract initial specific scientific data types. Subsequently, the preliminary extraction results are filtered through a list of candidate words to obtain the final results.

The prompt is as follows:

```
1 You need to identify and list any physical chemistry scientific quantities mentioned
  in the following experimental procedure.
2 Requirements:
3 1. The output format should be only one per line as "Original Text: Scientific
  Quantity".
4 2. Please extract information from the given sentences without creating your own
  summaries.
5 Output Example:
6 50mL: Volume
7 {}
8 The given experimental procedure are:
9 {}
```

The filtering word list is as follows:

```
1 "Volume", "Temperature", "Length",
2 "Energy", "Concentration", "Mass", "Speed", "Acceleration", "Density", "Frequency",
3 "Force", "Acidity", "Flow Rate", "Pressure", "Voltage"
```

### C.3 Implementation of feature vector transformation

Each  $\mathbf{x}_i$  in the dataset  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  is a feature vector representing an operation pattern extracted from the corpora. To obtain this dataset, we extract all sentences starting with verbs from the dataset using NLTK's part-of-speech tagging firstly. We then utilize GPT to annotate the parameters within these sentences and ultimately convert them into ont-hot feature vectors.

The prompt is as follows:

```
1 You need to complete an entity recognition task with defined entity categories: {}
2 Requirements:
3 1. The output format should be annotated on the original sentence, and only the
  annotated sentence should be output.
4 2. Please extract information from the given sentences without creating your own
  summaries.
5 3. Text is in a laboratory setting, please carefully analyze the specialized terms
  in the fields of biology and chemistry.
6 4. Please extract as many entities as possible from this sentence.
7 Example:
8 {}
9 The given sentences are:
10 {}
```

### C.4 Implementation of DPMM

**Spectral clustering via DPMM** The DPMM facilitates the spectral clustering of these feature vectors into groups representing unique atomic semantic constraints, accommodating the variability and complexity of semantic patterns across different domain corpora. The iterative clustering process, centered around Gibbs sampling, estimates the posterior distributions of the model's parameters. This not only identifies distinct semantic constraints but also adapts to the intricate semantic patterns prevalent in various domain-specified corpora.

**Algorithm symbols and initialization** Key symbols include  $\mathcal{X}$  for the dataset of feature vectors;  $N$ , the total number of samples; and  $L$ , the flexible number of clusters within the mixture model. Latent variables  $\mathcal{W}$  indicate cluster assignments for each  $\mathbf{x}_i$ , while  $\Psi$  encompasses the set of parameters for each



cluster. The concentration parameter  $\alpha$  influences new cluster formation likelihood, with  $\sigma_m$  serving as the regularization term for parameter updates.

Initially,  $K = 1$  and cluster parameters  $\psi_0$  are set with generic values, *e.g.*,  $\mu = 0$ ,  $\sigma = 1$ , assigning all data points to an initial cluster ( $w_i = 0$  for each  $i$ ).

**Gibbs sampling and posterior inference** The core of the algorithm lies in the Gibbs sampling loop, which iterates until convergence. It updates the cluster assignment  $w_i$  for each  $\mathbf{x}_i$  by evaluating the posterior distribution that incorporates the current parameter estimates and  $\alpha$ . The posterior probability of  $\mathbf{x}_i$  belonging to cluster  $l$  integrates the likelihood of observing  $\mathbf{x}_i$  given the cluster parameters and the prior probability of cluster membership, formalized as:

$$p(w_i = l | \mathbf{w}_{-i}, \mathbf{x}_i, \alpha, \Psi) \propto p(\mathbf{x}_i | \psi_l) \cdot p(w_i = l | \mathbf{w}_{-i}, \alpha), \quad (\text{A1})$$

where  $p(\mathbf{x}_i | \psi_l)$  is the likelihood of observing  $\mathbf{x}_i$  under the parameters of cluster  $l$ , and  $p(w_i = l | \mathbf{w}_{-i}, \alpha)$  reflects the adjusted prior probability for existing clusters and the potential for new cluster formation, influenced by  $\alpha$ .

Following the assignment update, cluster parameters  $\Psi$  are recalibrated using Maximum A Posteriori (MAP), applying regularization with  $\sigma_m$  to ensure stability and prevent overfitting.

## C.5 Exemplar output

There are several examples demonstrating the abstraction from natural language instructions to semantic patterns, from domains of Medical, Genetics, and Ecology.

Medical:

```

1 {
2   "TRANSECT": [
3     {
4       "pattern": ["REG", "REG"],
5       "example": [
6         "Transect the [aorta]{REG} proximally to the origin of the [
brachiocephalic trunk]{REG}.",
7         "Transect the [bile duct]{REG} close to the [pancreas]{REG}.",
8         "Transect the [SHVC]{REG} along with part of the diaphragm and [IHVC
]{REG} at the level of the left renal vein when the donor liver became pallid.",
9         "Transect the liver parenchyma of the [caudate lobe]{REG} and the [
Spiegel lobe]{REG}.",
10        "Transect the [infrahepatic inferior vena cava (IHIVC)]{REG} and
mobilize the [LVC]{REG} from retroperitoneal tissue."
11      ],
12    },
13    {
14      "pattern": ["REG", "Device"],
15      "example": [
16        "Transect the [cranial nerves]{REG} with the [scissors]{Device}.",
17        "Transect the [PHA]{REG} of the recipient at its root to expose the
vascular lumen using [micro-scissors]{Device}.",
18        "Transect the [femoral artery]{REG} in the section located between
the distal and proximal knots using [spring scissors]{Device}."
19      ],
20    },
21    {
22      "pattern": ["REG", "Device", "REG"],
23      "example": [
24        "Transect the [aorta]{REG} with [fine sharp scissors]{Device} (see
Table of Materials) just proximal to the [subclavian artery]{REG} takeoff."
25      ],

```

```

26     },
27     {
28         "pattern": ["REG"],
29         "example": [
30             "Transect the [colon]{REG}.",
31             "Transect the [umbilical ligament]{REG}.",
32             "Transect the [IVC]{REG} 1 cm above the diaphragm.",
33             "Transect the [stretcher/opener motor nerve]{REG}."
34         ],
35     }
36 ]
37 }

```

### Genetics:

```

1 {
2     "RESOLVE": [
3         {
4             "pattern": ["REG", "REG", "REG", "Length"],
5             "example": [
6                 "Resolved 1 \u00b5l of the [PCR reaction]{REG} on a [DNA gel]{REG}
to confirm the successful reaction with a [DNA band]{REG} at about [10 kb]{
Length}.",
7                 "Resolve 2 \u00b5l of the [PCR reaction]{REG} on a [DNA gel]{REG} to
confirm the successful reaction to see if a [PCR product]{REG} of [3.5 kb]{
Length} is present."
8             ],
9         },
10        {
11            "pattern": ["Concentration", "Concentration"],
12            "example": [
13                "Resolve [0.05%]{Concentration} of the input (=extract/lysate) and
at least [10%]{Concentration} of the eluates on a 4\u201315% Mini-PROTEAN\u00ae
TGX\u2122 Precast SDS-PAA gel."
14            ],
15        },
16        {
17            "pattern": ["REG", "REG"],
18            "example": [
19                "Resolve the cell pellet in [PBS]{REG} and [centrifuge]{REG} again
.",
20                "Resolve [5 \u00b5L]{Volume} of the [PCR]{REG} from [6-12]{String}
representative samples on a [2% (w/v)]{Concentration} [agarose gel]{REG}.",
21                "Resolve [10 \u00b5l]{Volume} of [PCR reactions]{REG} on [1%]{
Concentration} [agarose gel]{Container} ([TAE]{REG} or [TBE]{REG}) with a [1 kb
]{Length} [DNA ladder]{REG} to check for positive amplification ([Figure 3C]{
String})."
22            ]
23        }
24    ]
25 }

```

### Ecology:

```

1 {
2     "DRILL": [
3         {
4             "pattern": ["Device"],

```

```

5     "example": [
6         "Drill a hole big enough to allow the [sensor]{Device} to be
inserted.",
7         "Drill a hole at this coordinate using a [medium tip burr]{Device} (
Figure 3).",
8     ],
9     },
10    {
11        "pattern": [
12            "Length"
13        ],
14        "example": [
15            "Drill a hole (approximately [7 mm]{Length}) in the skull above the
target area using coordinates from a brain atlas according to the animal used.",
16            "Drill a craniotomy of approximately [1 mm]{Length} in diameter in
this location.",
17            "Drill the [craniotomy]{Length}."
18        ],
19    },
20    {
21        "pattern": [
22            "REG"
23        ],
24        "example": [
25            "Drill out the hole in the [pulley]{REG} to match the diameter of
the [motor shaft]{REG}.",
26            "Drill the [skull]{REG} on the marked point using a .9 mm diameter
steel drill tip."
27        ],
28    },
29    {
30        "pattern": [
31            "Length",
32            "Device"
33        ],
34        "example": [
35            "Drill a [3 in]{Length} hole in the top panel for the [fan]{Device
}."",
36            "Drill a small burr hole (approximately [0.5 mm]{Length}) using the
[Micro-Drill]{Device} at the [coordinates]{REG}."
37        ],
38    }
39 ],
40 }

```

## C.6 Extended discussions

**On the assignments of  $K$  and  $K'$**  Empirically,  $K'$  and  $K$  are in the same order of magnitude, with a difference roughly ranging in 25% ~ 35%, in the experiments we have conducted. Intuitively,  $K'$  denotes an estimation of syntactic constraints before optimization and  $K$  is the exact total number of syntactic constraints after optimization.

$K'$  is specified by an appropriate initialization. It is grounded to a predefined upper bound according to the size of the general syntax tree for C/C++. During the optimization process, the syntax tree is pruned according to the observed distribution of syntactic constraints on the corresponding domain corpus.

**On the scale of training corpus** In our studies, we have found that approximately 348 protocols, with an average of 121 steps per protocol, totaling 42,108 steps, are needed as a minimum amount of domain-specific corpus for the AutoDSL framework to work on a specific domain and to design a DSL accordingly. This threshold ensures that the dataset possesses sufficient richness of domain-specific context and is robust for extracting syntax and semantic constraints.

## D Data sources

### D.1 Corpora

Our corpora come from five websites including Nature<sup>2</sup>, Cell<sup>3</sup>, Bio<sup>4</sup>, Wiley<sup>5</sup> and Jove<sup>6</sup>. From a domain perspective, our corpora are classified into 5 domains including BioEng, InfoBio, Medical, Ecology and Genetics. We quantified each domain corpus in Table A2 by assessing the website sources, total number of protocols, and the average number of steps, actions, and objects per protocol. We only use sentences with subject-verb-object and predicate-object constructions to count the steps, actions, and objects. In addition, we performed the part-of-speech tagging of each word in the sentence, treating the verb as an action and the noun as an object. To note, we take the prototype of the verb or noun for the statistics.

Table A2: Statistics of the corpora across five domains

| Domain          | Nature | Cell | Bio  | Wiley | Jove | total | step | action type | object type |
|-----------------|--------|------|------|-------|------|-------|------|-------------|-------------|
| <b>InfoBio</b>  | 0      | 298  | 0    | 50    | 0    | 348   | 121  | 120         | 280         |
| <b>Ecology</b>  | 55     | 166  | 0    | 34    | 557  | 812   | 71   | 88          | 193         |
| <b>BioEng</b>   | 941    | 1404 | 0    | 22    | 1230 | 3597  | 70   | 80          | 188         |
| <b>Medical</b>  | 290    | 1530 | 1061 | 116   | 4354 | 7351  | 65   | 80          | 189         |
| <b>Genetics</b> | 1045   | 2296 | 3522 | 134   | 1797 | 8794  | 58   | 73          | 175         |

We demonstrate procedure cases from five domains with respect to sources.

```

1 [BioEng/Nature]
2 1. Virulent S. aureus MB 2865 \ (Smith strain) was grown overnight with aeration in
   TSB medium.
3 2. Bacteria were harvested and washed with fresh TSB and serially diluted to achieve
   an infective inoculum of 4 x 10<sup>3</sup> CFU/mL in 5% hog gastric mucin, and
   0.5 mL was used to infect intraperitoneally.
4 3. The infected mice were anesthetized using isoflurane \ (Abbott). 0.5 inch hollow
   connectors were attached to the external section of the JVC catheter and the
   mice were subsequently harnessed to the counter-balance arm inside a mouse cage.
5 4. Fluid primed PE 20 tubings attached to individual syringes on the infusion pump
   were connected to the mice through the 0.5 inch hollow connector.
6 5. The infusion pump was initiated for 24 hours at a flow rate of 0.1 mL/hour.
7 6. Blood samples are taken by tail vein to determine serum drug concentration during
   the infusion.
8 7. After the 24 hour period, the mice were euthanized and the kidneys were
   aseptically removed and homogenized.
9 8. Serial dilutions of the homogenates were plated on Mannitol plates and incubated
   overnight at 37 °C.
10 9. Bacterial counts were enumerated \ (ref. 2).

```

```

1 [InfoBio/Cell]
2 Timing: 1 h per experiment (for step 3)
3 The copulation-triggered photostimulation system detects copulation events

```

<sup>2</sup><https://protocolexchange.researchsquare.com/>

<sup>3</sup><https://star-protocols.cell.com/>

<sup>4</sup><https://bio-protocol.org/en>

<sup>5</sup><https://currentprotocols.onlinelibrary.wiley.com/>

<sup>6</sup><https://www.jove.com/>

4 in real time and turns on the green laser required for GtACR1-mediated  
5 inhibition of neural activity. This system allows for the inhibition of  
6 specific neural circuits in males only during copulation. In this assay,  
7 we use males that express GtACR111[[href=https://www.wicell.org#bib10](https://www.wicell.org#bib10)] in  
8 piezo-expressing neurons (w1118; +/+; piezo-GAL4/UAS-GtACR1.d.EYFP). As a  
9 mating partner, we use  
10 wild-type females. See  
11 key resources table[[href=https://www.wicell.org#key-resources-table](https://www.wicell.org#key-resources-table)] for  
12 details on fly  
13 strains.  
14 Prepare GtACR1-expressing male flies.  
15 Set up the fly crosses. Introduce parental flies (4-5  
16 piezo-GAL4 males and UAS-GtACR1 virgin females) into  
17 vials containing fly media. The detailed genotypes are listed in the  
18 key resources table[[href=https://www.wicell.org#key-resources-table](https://www.wicell.org#key-resources-table)].  
19 Transfer parental flies (F0 fly) to fresh vials every 5 days.  
20 Collect F1 males of the desired genotypes according to step 1 in  
21 before you begin[[href=https://www.wicell.org#before-you-begin](https://www.wicell.org#before-you-begin)].  
22 Note: F0 flies are transferred to fresh  
23 vials up to 2 times.  
24  
25 Note: Set-up several vials at a time to  
26 obtain enough F1 male flies.  
27  
28 Note: Wild-type females are prepared as  
29 mating partners according to step 1 in Before you begin.  
30  
31 Pause point: The experimenter can  
32 decide the timing of subsequent feeding of all-trans-retinal (ATR) as long  
33 as the tested flies are within 5-8 days after eclosion at the time of the  
34 optogenetic assay.  
35 Feeding of ATR. Male flies of the experimental group are fed food  
36 containing ATR. Control males are fed food containing ethanol (EtOH)  
37 solvent.  
38 Prepare plastic tubes containing ~150  $\mu$ L of fly food.  
39  
40 For the experimental group, add 2  $\mu$ L of ATR, 25 mg/mL dissolved in  
41 99.5% EtOH (need to be prepared just before use) on the food  
42 surface. For the control group, add 2  $\mu$ L of 99.5% EtOH on the food  
43 surface instead of ATR.  
44 Transfer males to a plastic tube containing fly food with the  
45 surface covered with ATR (or EtOH) 24-36 h before the experiment.  
46 Keep the male flies on prepared food for 24-36 h in the dark before  
47 being used for the assays.

1 [Medical/Bio]

2 Spleens were obtained from naive and malaria-infected C57BL/6 mice and were  
homogenized at room temperature in 6-well plates with 6 ml of HBSS with 2% FCS  
through a 70  $\mu$ m cell strainer to form single-cell suspensions. Red blood cells  
were lysed using 1 to 2 ml of RBC lysing buffer (depending on size of spleen)  
and splenocytes were washed once at 200 x g 4 °C for 5 min with cold FACS buffer  
(HBSS with 2% FCS). Splenocytes were resuspended by gentle tapping on a rack  
in FACS buffer and kept on ice at all times to avoid background phosphorylation  
of STAT proteins. Viability and cell counts were obtained by trypan blue

exclusion using a haemocytometer. Cells were washed once with 1 ml AIM V® Medium, resuspended at 20 x 10<sup>6</sup> cells/ml and rested on ice for a minimum of 20 min. 1 x 10<sup>6</sup> cells were incubated with 20 ng/ml rIL-2 or 2.5 ng/ml rIL-12 for 10 min at 37 °C, 5% CO<sub>2</sub> (final volume of 200 µl) and immediately fixed on ice for 15 min by adding an equal volume of 4% paraformaldehyde. Cells were washed with FACS buffer, resuspended in 500 µl of 90% ice-cold methanol and immediately stored down at -20 °C for a minimum of 2 h (cells can be kept for up to a month without affecting further staining). Splenocytes were washed twice with FACS buffer and stained in FACS buffer at room temperature for 30 min for CD4 (GK1.5), CD44 (IM7), CD62L (MEL-14), T-bet (4B10) and phosphorylated STAT4 (at residue Y693, clone 38) or phosphorylated STAT5 (at residue Y694, clone 47). Cells were washed with FACS buffer and analysed by flow cytometry.

Fluorescence minus one controls were included to validate flow cytometric results. Flow cytometry acquisition was performed using an LSR II.

1 [Ecology/Wiley]

2 Generating primary transformants

3 1. Sow seeds of the transformation on GM medium containing gentamicin to select for transgenic plants harboring the T-DNA.

4 2. Let the seeds mature for 2 weeks at 22°C in a growth chamber.

5 3. For GT: Pick at least 40 T1 plants containing pDe-EC-ttLbCas12a and transfer them to soil until maturity.

6 4. For cleavage activity: Extract DNA from 20 primary transformants containing the pDe-ttLbCas12a construct via the rapid DNA extraction method (see Edwards et al., 1991[<https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cppb.20117#cppb20117-bib-0002>]). Analyze the mutagenesis efficiency by e.g., TIDE (see Critical Parameters).

7 5. Extract DNA from one leaf of each plant and set up a suitable PCR to verify the presence of your construct.

8 Obtaining heritable GT plants

9 6. Harvest the seeds of each T1 plant separately.

10 7. Sow about 100 seeds per line on GM medium.

11 8. Let the seeds mature for 2 weeks at 22°C.

12 9. Extract DNA from one leaf of 100 plants per line as a pool.

13 10. Screen for positive GT events using a suitable PCR (see Critical Parameters).

14 11. Extract DNA from each plant of the positive identified T1 pools separately after another week of growth.

15 12. Analyze the T2 plants separately for heritable GT via PCR and confirm it by sequencing (see Critical Parameters).

16 13. Transfer positive T2 plants to soil and cultivate them to maturity.

1 [Genetics/Jove]

2 All experiments involving the differentiation of human iPSC lines were performed in compliance with the Institutional Review Board of Boston University (protocol H33122). The dermal fibroblasts, procured for reprogramming to iPSCs, were obtained from a donor with written informed consent, under the approval of the Human Research Protection Office of Washington University School of Medicine, St. Louis, MO. Reprogrammed iPSCs were generated at the Center for Regenerative Medicine at Boston University and Boston Medical Center, Boston, MA.

3 1. Alveolosphere dissociation

4 Prepare complete serum-free differentiation media (cSFDM) as per the composition mentioned in Table 1.

5 Prepare CK + DCI media in the prepared cSFDM base as per Table 2.

6 Thaw 2D (human embryonic stem cell-qualified) and/or 3D (growth-factor reduced) matrix on ice as required for the experimental needs.



- 7 Aspirate all the CK + DCI medium using a pipette or aspirating pipette with vacuum from the 3D matrix droplets containing alveolospheres, derived from directed differentiation<sup>19</sup>, in a 12-well plate.
- 8 Add 1 mL of dispase (2 mg/mL) per droplet. Gently pipette the droplet into the dispase using a P1000 pipette. Incubate at 37 °C for 1 h, pipetting up and down once after 30 min.
- 9 Transfer the dissociated organoids (from Step 1.5) from one matrix droplet in the dispase to a 15 mL conical tube. To wash, add 10 mL of Iscove's Modified Dulbecco's Medium (IMDM, see Table of Materials).
- 10 Centrifuge at 300 x g for 5 min at room temperature. Aspirate the supernatant using a pipette or aspirating pipette with vacuum, leaving as little supernatant as possible.
- 11 NOTE: It is important to remove all dispase as any remaining dispase may dissolve the matrix that the cells will subsequently be seeded into. If a clear haze is seen above the pellet, the dispase has not completely dissolved the matrix, and more dispase can be added to the pellet for another 20-30 min at 37 °C.
- 12 Resuspend the cells in 1 mL of 0.05% trypsin per droplet and transfer back to the 12-well plate. Incubate at 37 °C for 12-15 min. Observe the dissociation under a microscope. Avoid over-pipetting the cells at this stage.
- 13 NOTE: At the end of incubation, the cells need to achieve a single-cell suspension after pipetting 3-5 times with a P1000 pipette. For passaging iAT2s to ALI (Step 3), the trypsinization time needs to be minimized (maximum 12 min), such that the cells are in 2-3-cell clumps rather than single-cell suspension when ready for plating onto the cell culture insert.
- 14 Stop the action of trypsin with an equal volume of FBS-containing medium (10% ES-qualified FBS in DMEM). Centrifuge at 300 x g for 5 min at room temperature.
- 15 Wash the cells with 10 mL of IMDM. Centrifuge at 300 x g for 5 min at room temperature.
- 16 Resuspend the cells in an appropriate volume for counting, and then count the cells using a hemocytometer (see Table of Materials).
- 17 NOTE: From one confluent 50  $\mu$ L matrix droplet seeded at 400 cells/ $\mu$ L, the expected yield is 500,000 to  $1.5 \times 10^6$  cells per droplet.
- 18 Use the single-cell suspension of iAT2 cells to generate alveolospheres by plating in the 3D matrix (Step 2) and/or plating on cell culture inserts for ALI culture (Step 3).
- 19 2. 3D plating of iAT2s
- 20 After counting (Step 1.11), determine the number of desired cells to replate in the 3D matrix (400 cells/ $\mu$ L of the matrix with 50-100  $\mu$ L of 3D matrix droplets per well of a 12-well plate). Centrifuge the cells at 300 x g for 5 min at room temperature. Remove as much supernatant as possible using a pipette.
- 21 Resuspend the cells in the 3D matrix. Resuspend quickly and on ice, if needed, to prevent the matrix from polymerizing (which occurs when warm).
- 22 Use a P200 pipette to dispense one 3D matrix droplet per well into a pre-warmed 12-well plate. Pipette carefully to avoid creating bubbles in the matrix droplet. Do not allow the cell suspension to settle while dispensing multiple droplets.
- 23 Place the plate in a 37 °C incubator for 20-30 min to allow the matrix droplets to polymerize.
- 24 Add 1 mL of CK + DCI + 10  $\mu$ M of Y-27632 medium (see Table of Materials) per well to cover the matrix droplet.
- 25 After 72 h, change the medium to CK + DCI without 10  $\mu$ M of Y-27632.
- 26 Replace the medium with fresh CK + DCI every 48-72 h.
- 27 NOTE: iAT2s will typically need to be passaged approximately every 10-14 days, depending on cell line and plating density.

## D.2 Data preprocessing

We perform some preprocessing on the source data, including slice procedures and domain mapping.

**Slice procedures** We segment the procedures of the protocols without destroying their structure as much as possible. Specifically, we first split the procedures according to the regular delimiter '`[\n\r]+`' to obtain a series of sub-paragraphs. Then we merge the sub-paragraphs in order, and the token number of merged paragraphs does not exceed the predetermined maximum token (i.e., 300). If the number of tokens in the merged paragraph is never satisfied, we follow the separator '.?!' to split and re-merge the paragraphs at a fine-grained level.

**Domain mapping** The protocols from the five data sources have their own subject areas. We mapped the topic domains to the five domains developed by the experts according to the expert-set mapping table.

```
1 [Genetics]
2 Antibody
3 Biochemistry
4 Biomarkers
5 Biophysics
6 Cell Biology
7 Cell Differentiation
8 Cell Membrane
9 Chromatin Immunoprecipitation (ChIP)
10 Gene Expression
11 Genetics
12 Genomics
13 Human Genetics
14 Molecular Biology
15 Molecular/Chemical Probes
16 Mouse Biology
17 Protein Biochemistry
18 Protein Science
19 Proteomics
20 RNaseq
21 Sequence Analysis
22 Sequencing
23 Signal Transduction
24 Single Cell
25 Single-Molecule Assays
26 Structural Biology
27 Systems Biology
28 Microbiology
29 Developmental Biology
30 Model Organisms
31 -----
32 [Medical]
33 Cancer
34 Cancer Research
35 Cardiology
36 Diseases
37 Drug Discovery
38 Gastroenterology
39 Health Sciences
40 Immunology
41 In Situ Hybridization
42 Immunology and Infection
43 Medicine
44 Neuroscience
```

```

45 Oncology
46 Organoids
47 Physiology
48 Pharmacology
49 Psychology
50 Rheumatology
51 Stem Cells
52 Stem Cell Biology
53 Tissue Engineering
54 Toxicology
55 Urology
56 Health Humanities
57 -----
58 [Ecology]
59 Behavior
60 Ecology
61 Plant Sciences
62 Plant Biology
63 -----
64 [Bioeng]
65 Bioengineering
66 Biotechnology
67 Cell Culture
68 Microscopy
69 Biological Techniques
70 Cell Isolation
71 Cell Separation/Fractionation
72 Cell-Based Assays
73 Chemical Biology
74 Cytometry
75 Mass Cytometry
76 Protein Expression And Purification
77 -----
78 [Infobio]
79 Bioinformatics
80 Computational Biology and Bioinformatics

```

## E Constraint design assessment details

### E.1 Meta evaluation

For meta evaluation, the experts' participation will include completing a survey and possibly partaking in an interview. The survey will present a series of statements or questions about the DSL. For each item, we request two responses from the experts: Opinion on a 1-5 Scale and Confidence in the Response on a 1-5 Scale. This dual-scale approach will help quantify both the perspective on the DSL and confidence in each response, providing a richer dataset for analysis.

In this part, we organized three experts in the field of experimental sciences to independently score our DSL. Each expert scored the results of a single DSL output across 9 dimensions, with a total of 50 questions. This process will be repeated 5 times, with experts independently evaluating 5 different subsets. We collected the scores from the three experts for subsequent processing and comparison.

The participants are instructed by the following guidance:

```

1 Invitation to Participate in the Evaluation of a New Domain-Specific Language for
  Experimental Sciences
2
3 Dear Expert,

```

4  
5 We are reaching out to invite you to partake in a crucial evaluation of a newly  
developed Domain-Specific Language (DSL) designed for life sciences. Your  
expertise in this field is invaluable for this assessment.

6  
7 Purpose of the Experiment:  
8 Our team has developed a DSL to facilitate both computerized and manual  
experimentation processes in life sciences. This evaluation aims to gather  
detailed feedback from experts like you to refine the DSL's usability and  
functionality.

9  
10 What Will Be Involved:  
11 Your participation will include completing a survey and possibly partaking in an  
interview. The survey will present a series of statements or questions about the  
DSL. For each item, we request two responses from you:

12 1. Opinion on a 1-5 Scale: Rate each item on a scale where 1 indicates 'Strongly  
Disagree' and 5 indicates 'Strongly Agree'. The scale is nuanced as follows:

13 - 1: Strong Disagreement or Major Issues  
14 - 2: Disagreement or Notable Concerns  
15 - 3: Neutral or Mixed Feelings  
16 - 4: Agreement or Minor Concerns  
17 - 5: Strong Agreement or Highly Favorable

18 2. Confidence in Your Response on a 1-5 Scale: Indicate your level of confidence  
in your response to each item, where 1 is 'Not Confident at All' and 5 is '  
Extremely Confident'. The scale implies:

19 - 1: Very Low Confidence  
20 - 2: Low Confidence  
21 - 3: Moderate Confidence  
22 - 4: High Confidence  
23 - 5: Very High Confidence

24  
25 This dual-scale approach will help quantify both your perspective on the DSL and  
your confidence in each response, providing a richer dataset for analysis.

26  
27 In the interview, we will delve deeper into your experiences with the DSL, allowing  
you to share more comprehensive insights and suggestions.

28  
29 Estimated Time Commitment:  
30 The survey should take approximately 60 minutes to complete. The interview, if you  
choose to participate, will be a 1-hour session, scheduled at your convenience.

31  
32 Confidentiality and Use of Data:  
33 Your responses will be kept strictly confidential and will be used solely for  
improving the DSL. We adhere to the highest ethical standards in our research.

34  
35 Your expertise and nuanced feedback are vital for the success of this project, and  
we greatly appreciate your consideration.

36  
37 Thank you for your time and expertise.

## E.2 Empirical evaluation

We take the GPT-4 as our expert in machine empirical evaluation. We built a knowledge augmentation system using Langchain<sup>7</sup> to provide additional domain knowledge to GPT-4. We designed our instruction

<sup>7</sup><https://www.langchain.com/>

template following the prompt format of Zheng et al. (2023) and utilizing the Plan-and-Solve decomposition idea (Wang et al., 2023). The machine evaluation results on DSL of five domains and BioCoder are shown in Table A3.

Table A3: Empirical evaluation by the QA System expert on the five DSLs corresponding to five domains and BioCoder

| Domain          | Opinion | Confidence |
|-----------------|---------|------------|
| <b>BioCoder</b> | 4.35    | 3.96       |
| <b>InfoBio</b>  | 4.38    | 3.97       |
| <b>Ecology</b>  | 4.39    | 3.96       |
| <b>BioEng</b>   | 4.41    | 3.97       |
| <b>Medical</b>  | 4.40    | 3.97       |
| <b>Genetics</b> | 4.41    | 3.97       |

The instruction template is displayed as follows:

```

1 Please act as an impartial judge and evaluate the given developed Domain-Specific
  Language ([DSL]) according to the [Rating Basis], [Rating Criteria], and related
  [Domain Knowledges]. Please follow the [Instruction]. Moreover, the [Background
  ] describes something about Domain-Specific Language ([DSL]).
2 -----
3 [Background]
4 Domain experts expect to describe the experimental protocols of the specific domain
  with the \ac{dsl} programs.
5 Domain experts expect to be guided by the \ac{dsl} programs step-by-step, where the
  input, output, and configurations of each step is well-detailed.
6 Domain experts come without any training on programming.
7 Domain experts are subjective and their mindsets come from the specific domain.
8
9 [Instruction]
10 The [Rating Basis] presents a series of statements or questions about the [DSL].
  Please evaluate your opinion and confidence level in the statements or questions
  referring to related domain knowledge.
11 That is, your evaluation should be based primarily [Rating Basis]. You should refer
  to and use [Domain Knowledge] to help you evaluate. Begin your evaluation by
  providing a short explanation. After providing your explanation, please rate the
  [DSL] on [Rating Criteria] by strictly following this format: "Rating: [[<
  opinion>, <confidence>]]", for example: "Rating: [[3, 4]]".
12
13 [Rating Criteria]
14 For each item, you need give an <Opinion> and <Confidence> rate:
15 Opinion on a 1-5 Scale (int): Rate each item on a scale where 1 indicates 'Strongly
  Disagree' and 5 indicates 'Strongly Agree'. The scale is nuanced as follows:
16 1: Strong Disagreement or Major Issues
17 2: Disagreement or Notable Concerns
18 3: Neutral or Mixed Feelings
19 4: Agreement or Minor Concerns
20 5: Strong Agreement or Highly Favorable
21 Confidence in Your Response on a 1-5 Scale (int): Indicate your level of confidence
  in your response to each item, where 1 is 'Not Confident at All' and 5 is '
  Extremely Confident'. The scale implies:
22 1: Very Low Confidence
23 2: Low Confidence
24 3: Moderate Confidence
25 4: High Confidence
26 5: Very High Confidence
27

```

```
28 [Rating Basis]
29 {}
30
31 [DSL]
32 {}
33
34 [Domain Knowledges]
35 {}
36
37 [Output]
38 Let's first understand/follow the [Instruction] to evaluate the [DSL] and give an
   explanation about rating.
39 Then let's give a final rating based on the explanation.
```

[Rating Basis] refers to our designed questions or statements.

```
1 "The DSL clearly represents the essential concepts necessary for life sciences
   experiments."
2 "All domain concepts included in the DSL are relevant and contribute directly to my
   tasks in life sciences."
3 "I can easily express all necessary domain concepts for my experiments using the DSL
   ."
4 "The DSL does not include complex concepts that are unnecessary for my work in life
   sciences."
5 "The DSL seems to have been developed with practical feedback from domain experts,
   ensuring its relevance to my work."
6 "The design of the DSL focuses on the tasks I need to accomplish in life sciences,
   avoiding irrelevant features."
7 "The DSL is simple and straightforward to understand and use in my life sciences
   work."
8 "Learning to use the DSL does not require extensive time or effort, even for those
   new to this kind of language."
9 "The tools and features of the DSL are clear and intuitive to use in my everyday
   tasks."
10 "The simplicity of the DSL enhances my productivity in conducting life sciences
   experiments."
11 "The straightforward nature of the DSL lowers the barrier to its adoption in my
   professional environment."
12 "The DSL specifically addresses the unique concepts and needs of my work in life
   sciences, without unnecessary generalizations."
13 "All concepts and features in the DSL are directly applicable and useful for my life
   sciences tasks."
14 "The DSL avoids overly complex or overly generalized concepts that is uneasy for me
   to understand, making it more practical for my work."
15 "The focused nature of the DSL facilitates its quick and successful introduction
   into my life sciences work."
16 "The DSL is concise and precise, focusing only on what is necessary for life
   sciences, without extraneous features."
17 "The DSL effectively meets the current needs of my work in life sciences without
   complicating future changes in the domain."
18 "The number of elements in the DSL is appropriate and manageable for my work in life
   sciences, making it easy to understand."
19 "The DSL effectively uses specialized sublanguages (or subsets) for different
   aspects of life sciences, enhancing clarity and focus."
20 "Despite having a limited number of elements, the DSL is effective in handling the
   complex tasks of my life sciences work."
```



21 "The DSL's approach to representing different aspects of life sciences work (like structure, behavior, etc.) is clear and well-organized."

22 "The DSL's limited elements are versatile enough to be applicable across a wide range of life sciences applications."

23 "Each concept in the DSL is distinct and clearly defined, with no overlap or redundancy in their functions or purposes."

24 "I find it straightforward to express ideas in my life sciences work using the DSL, as it avoids unnecessary duplication of concepts."

25 "The DSL allows for consistent modeling of facts and processes in life sciences, without confusion due to redundant concepts."

26 "Remembering and applying different concepts in the DSL is easy, as it avoids minor variations that could lead to confusion."

27 "The DSL effectively represents life sciences data and processes without complicating them through redundant concepts."

28 "The non-redundant design of the DSL positively impacts my workflow in life sciences by providing clarity and efficiency."

29 "The notations used in the DSL are familiar to me, reflecting those commonly used in my life sciences work."

30 "Transitioning to using the DSL was easy for me because it adopts notations and terminologies I'm already accustomed to in life sciences."

31 "I did not have to spend much time learning new syntaxes or notations when starting with the DSL, thanks to its use of familiar concepts."

32 "The syntax and notation of the DSL are well-suited for the specific tasks I perform in life sciences."

33 "The DSL effectively integrates languages or notations that are commonly accepted and used in my field (like SQL for database queries, if relevant)."

34 "The DSL aligns closely with the existing notations and terminologies that are standard in the life sciences domain."

35 "In the DSL, different elements are clearly distinguishable from one another, enhancing my understanding and ease of use."

36 "In textual aspects of the DSL, keywords and language elements are positioned in a way that makes the content easy to read and understand."

37 "The DSL effectively avoids ambiguity in representing different domain concepts, making it simple for me to interpret the models or text."

38 "The design of the DSL prioritizes readability and understanding for the reader, rather than writing efficiency."

39 "The DSL allows me to easily add comments to model elements, aiding in documentation and explanation."

40 "Comments within the DSL models significantly enhance the understandability and clarity of my work."

41 "The ability to comment on model elements in the DSL simplifies and supports collaborative efforts in my life sciences projects."

42 "The DSL offers flexibility in commenting styles, such as line comments and block comments for textual languages or annotations for graphical elements."

43 "Comments in the DSL can be effectively used for detailed documentation, similar to generating HTML pages or Javadoc."

44 "The DSL encourages or enforces a consistent style of documentation through its comment structure, improving overall model quality."

45 "The usage conventions defined for the DSL contribute to its clarity and comprehensibility, making it easier for me to use in my life sciences work."

46 "The DSL strikes a good balance between providing guidance through usage conventions and allowing flexibility in how I use it."

47 "The established conventions in the DSL aid in maintaining the quality and consistency of my work."

48 "Following the usage conventions in the DSL is straightforward and does not add

unnecessary complexity to my tasks."  
 49 "The conventions in the DSL, such as the notation of identifiers and the order of  
 elements, enhance the overall comprehensibility of the language."  
 50 "The usage conventions of the DSL positively impact its overall usability and  
 practicality in my life sciences work."

**Rationale of scoring** We showcase some non-cherry-picked scoring rationales given by both human expert and our QA System expert in.

Table A4: **Showcases of scoring rationales.**(the corresponding scoring is attached [**Opinion, Confidence**])

| Criteria   | human expert   | QA System expert   |
|--|--|--|
| "The DSL clearly represents the essential concepts necessary for life sciences experiments."                                   | The DSL precisely covers key life sciences experiments, enhancing protocol clarity and accuracy. [5, 4]  | The DSL for life sciences effectively encapsulates key experimental actions—ASPIRATE, ADD, CLEAN, COMBINE, PUT—covering essential concepts from aspirating liquids to placing items. Its design and domain-specific actions provide a comprehensive framework for describing experiments, reflecting its adequacy for the life sciences domain. [5, 4] |
| "The DSL allows for consistent modeling of facts and processes in life sciences, without confusion due to redundant concepts." | The DSL clarifies life sciences protocols, potentially outperforming traditional methods with its precision. [4, 4]  | The DSL effectively captures key experimental actions in life sciences, offering clarity with specific patterns and examples for each action. Rated with high confidence, it ensures precise protocol definition, minimizing confusion and redundancy. Further analysis could enhance its validation against existing standards. [4, 4]                |
| "The DSL is simple and straightforward to understand and use in my life sciences work."  | The DSL simplifies life sciences protocols with clear commands and examples, though it may benefit from expanded flexibility. [4, 4]                       | The DSL is effectively designed for life sciences, offering domain-specific commands like ASPIRATE and ADD with clear examples, facilitating experimental protocols without programming knowledge. While comprehensive and user-friendly, potential areas for additional commands or flexibility could further enhance its utility. [4, 4]             |
| "The DSL's limited elements are versatile enough to be applicable across a wide range of life sciences applications."          | The DSL streamlines essential life sciences experiments but may fall short in specialized contexts, justifying a cautious yet positive rating of 4. [4, 4] | The DSL effectively guides life sciences experts through standard experiments with commands like ASPIRATE and ADD, balancing simplicity and specificity. Rated at 4, it offers a solid foundation for common protocols but may need expansion for more specialized experiments. [4, 4]   |

### E.3 Ontology extraction

We take a zero-shot IE tool as our backbone to extract ontologies from protocols. We add the domain-related definition to the Stage-I prompt of RE task in ChatIE, which is defined as follows: ...In this task, an entity type may be a 'OpCode' (operations, a one-word verb, like ADD and REMOVE); 'REG' (reagents taking part in an operation, like cells and MDDC culture media); 'COND' (conditions of executing an operation, like <temperature> 37C, <time> 30min, <device> a small scissor and <container> PCR tubes)....

In addition, we present the entity relation mapping dictionary as follows:

```
1 {
2 'is concurrent with': ['OpCode', 'OpCode'],
3 'is instruction of': ['OpCode', 'REG/COND'],
4 'is predecessor of': ['OpCode', 'OpCode'],
```

```

5 'is product of': ['REG', 'OpCode'],
6 'is reaction acceleration of': ['Acceleration', 'OpCode'],
7 'is reaction centrifugal force of': ['Centrifugal Force', 'OpCode'],
8 'is reaction condition of': ['COND', 'OpCode'],
9 'is reaction container of': ['Container', 'OpCode'],
10 'is reaction density of': ['Density', 'OpCode'],
11 'is reaction device of': ['Device', 'OpCode'],
12 'is reaction energy of': ['Energy', 'OpCode'],
13 'is reaction flow rate of': ['Flow Rate', 'OpCode'],
14 'is reaction force of': ['Force', 'OpCode'],
15 'is reaction frequency of': ['Frequency', 'OpCode'],
16 'is reaction iteration count of': ['Iteration Count', 'OpCode'],
17 'is reaction pressure of': ['Pressure', 'OpCode'],
18 'is reaction rotation of': ['Rotation', 'OpCode'],
19 'is reaction speed of': ['Speed', 'OpCode'],
20 'is reaction temperature of': ['Temperature', 'OpCode'],
21 'is reaction time of': ['Time', 'OpCode'],
22 'is reaction voltage of': ['Voltage', 'OpCode'],
23 'is reagent acidity of': ['Acidity', 'OpCode'],
24 'is reagent coating of': ['Coating', 'OpCode'],
25 'is reagent concentration of': ['Concentration', 'OpCode'],
26 'is reagent density of': ['Density', 'OpCode'],
27 'is reagent length of': ['Length', 'OpCode'],
28 'is reagent mass of': ['Mass', 'OpCode'],
29 'is reagent medium of': ['Medium', 'OpCode'],
30 'is reagent of': ['REG', 'OpCode'],
31 'is reagent quantity of': ['Quantity', 'OpCode'],
32 'is reagent size of': ['Size', 'OpCode'],
33 'is reagent thickness of': ['Thickness', 'OpCode'],
34 'is reagent volume of': ['Volume', 'OpCode'],
35 'is successor of': ['OpCode', 'OpCode']
36 }

```

## F Constraint utility assessment details

### F.1 Protocol source

In order to better analyze the subsequent DSLs, we created a protocol dataset of five domains in the life sciences in this paper, where all the protocols are not pre-existing protocols, but NOVEL ones, designed by ourselves, and of course most of these designed protocols are unpublished and written with reference to a large number of existing protocols.

### F.2 Protocol constraint

The protocol dataset we built covers both syntactic constraints and semantic constraints. We categorized these two main features into multiple types and built specific and typical protocols for each subclass, which can encompass most of the experimental scenarios.

**Syntactic constraint** Syntactic constraints can be primarily divided into four categories: the first category, the imperative model, includes loop, branch, jump, memory (allocation and de-allocation), function/(sub)procedure, and logical test. The second category, the type system, encompasses integer types, real/floating types, set types, scientific types, reagent types, and container types. The third category, concurrent, contains data parallel. The fourth category is the reactive model, which responds to a certain event.

```

1 {
2 Loop constraint
3 Cell culture:

```

```

4 1. Prepare medium: prepare appropriate medium according to cell type.
5 2. Inoculate: Add cell suspension to culture flasks containing culture medium.
6 3. Continuous monitoring: Observe the cell growth status and density periodically.
  Repeat the culture as long as the cells have not reached the harvested density.
7 4. Harvest cells: When the cells reach the expected density, harvest the cells for
  the next experiment or for passaging culture.
8 }
9 {
10 Branch constraint
11 Brown Adipocyte Respiration Protocol: Prepare cells as described in the first two
  steps of the intact brown adipocyte respiration protocol. On the day of the
  experiment, thaw 1x MAS and pre-made substrate solutions. Prepare 10x solutions
  of inhibitors in 1x MAS. If using fatty acid substrates, prepare them freshly as
  described. Soak the cartridge for at least 3 hours before loading. Load
  specified volumes into ports A-D and calibrate the cartridge. Prepare 7.5 nM PMP
  in 1x MAS, gently wash cells with 1x MAS once, completely evacuate the wash,
  and add 450  $\mu$ l 1x MAS with 7.5 nM PMP. Perform the Seahorse assay with 2-3
  measurements per condition, ensuring not to exceed 1 hour before injecting
  Antimycin A to maintain mitochondrial integrity.
12 }
13 {
14 Memory constraint
15 Bile Processing Protocol: Collect bile in sterile bottles from patients with biliary
  tubes. Centrifuge at 3000g for 10 minutes at 4°C to remove sediment and debris.
  Aliquot the clear bile into 1 ml volumes in microcentrifuge tubes. Store the
  aliquots at -20°C for future use.
16 }
17 {
18 Function procedure constraint
19 Isolation of G4 Deletion Alleles in C. elegans: Clone out dog-1 animals onto plates,
  optionally freeze half as backup. Rinse worms in M9, transfer to PCR plates,
  add lysis buffer, and proceed with lysis and proteinase K inactivation. Perform
  nested PCR on lysis mix to identify positive populations for germline mutations.
  Analyze PCR products on agarose gel. For positive populations, chunk
  corresponding plate, grow to starvation, and repeat the process to confirm
  deletions. Perform nested PCR on progeny to confirm homozygosity of deletion
  alleles. Sequence PCR product for deletion characterization. Backcross new
  strain to N2 to clean genetic background.
20 }
21 {
22 Parallel constraint
23 Centrifugation and RNA Isolation from Serum or Plasma: Start with 1 ml of serum or
  plasma for each replicate. First centrifugation: Bring volume up with PBS and
  centrifuge all samples at once for 90 min at 100,000 x g. Second centrifugation:
  After discarding supernatant and resuspending pellets, centrifuge again for 70
  min at 100,000 x g. Parallel processing: Process multiple samples through two
  rounds of centrifugation and subsequent RNA isolation, highlighting batch
  processing of samples.
24 }
25 {
26 Reactive model constraint
27 Cell Growth and Transfection Response: Thaw and culture cells until 80% confluency.
  Response to 80% confluency: Split cells for further growth or prepare for
  freezing. Transfect cells at low density, then change media based on purpose (
  growth or differentiation). Response to transfection: Monitor for complete

```

```
differentiation, adjusting care based on confluency or media type."
```

```
28 }
```

**Semantic constraint** Semantic constraints mainly include two types: (i) Use of undefined action; (ii) Incomplete parameter, where the first category is required parameters do not exist, and the second category is Required parameter under-specified (is not grounded to the granularity for execution).

### F.3 Prompt engineering for utility assessment

Following methods suggested by recent research (Gao et al., 2023; Zhang et al., 2023), we prompt the LLM with examples of DSL syntax and semantics, directing it to translate procedural texts into corresponding programs. To ensure a fair comparison and to underscore the plug-and-play capability of the DSLs designed by AutoDSL, only minimal prompt engineering is applied in protocol processing. The resulting prompt text under the ideas is demonstrated as follows.

```
1 You are an expert in life science and computer science. Now you are prompted with a
  grammar of programming language defined by production rules, several experiment
  steps described in natural language, and a construct (which is the left part in
  the production rule). Your task is to determine whether the natural language
  description consists of parts that can be parsed using this production rule. If
  the natural language string can be parsed, please output "Yes", otherwise,
  output "No".
2
3 The production rule:
4 for (initialization; condition; increment) statement
5
6 The nature language experimental protocol:
7 Study design: Timing: 3 days for protocol development. IRB/ethics review and
  revision can take 4-8 weeks. Repeat with number from 1 to 11:1. Define the
  crisis situation.2. Select existing algorithms for study.3. Define hypothetical
  triage algorithms for study.4. Select primary and secondary clinical endpoints
  .5. Determine the clinical data needed.6. Select the patient cohort for study.7.
  Assess availability of necessary clinical data.8. Adapt the triage algorithm
  scoring scheme.9. Submit the proposed study to the IRB or ethics panel.
  Calculating priority scores: Timing: 3 weeks to 3 months. Repeat with number
  from 1 to 8:1. Create a case report form.2. Select a data management system.3.
  Select the method of data entry. 4. Perform pilot data acquisition. 5. Complete
  the case report form for all patients.6. Apply the triage algorithms and
  calculate priority scores. Testing algorithm accuracy Timing: 1 day.1. Determine
  the accuracy of the priority scores Simulation of clinical decision-making:
  Selection from a smaller group Timing: 1 day. Repeat with number from 1 to 4:1.
  Format input files.2. Run CSC script.3. Assess bootstrap analysis output.
  Sensitivity analyses: Timing: 1 week. Repeat with number from 1 to 4:1. Test the
  effect of data processing methods.2. Test the effect of patient characteristics
  . 3. Test the effect of triage algorithm components.
8 yes
9
10 The production rule:
11 for (initialization; condition; increment) statement
12
13 The nature language experimental protocol:
14 Mitochondria Sonication Experiment:Repeat sonication of mitochondria in a Branson
  450 sonicator using aconitase buffer (50 mM Tris, 30 mM sodium citrate, 0.5 mM
  MnCl2, 0.2 mM NADP, pH 7.3) four times, each for 15 seconds.Monitor the citrate
  to alpha-ketoglutarate conversion at 340 nm at 25°C, utilizing 2 units/ml of
  isocitrate dehydrogenase in 50 mM Tris, 1 mM cysteine, 1 mM sodium citrate, 0.5
  mM MnCl2 at pH 7.4. Reactivate aconitase with 2 mM dithiothreitol and 0.2 mM
```

ferrous ammonium sulfate for 5 minutes, then repeat the enzymatic activity assay once.

15

16 Task:

17 Please use a JSON format to describe the protocol (Only output a json).

18

19 Note:

20 1. Read the protocol carefully.

21 2. Choose items and conditions from the protocol to describe the protocol.

22 3. Your desired output format resembles this: {"opcode": [['Datatype', 'Data'], ...  
 , ["output", "Data"]]. In this format, "Data" = None indicates a missing value.

23 4. You must specify the output in json, and only one output in json is allowed.

24

25 Example:

26 GROW:[['REG', 'GFP-fusion'], ["REG", "solid media"], ["Device", "spinning disk  
confocal microscope"]] -> Z-stack images

27 Output: {"GROW": [['REG', 'GFP-fusion'], ["REG", "solid media"], ["Device", "  
spinning disk confocal microscope"], ["output", "Z-stack images"]]}