DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines

Omar Khattab, Arnav Singhvi,
Paridhi Maheshwari, Zhiyuan Zhang,
Keshav Santhanam, Sri Vardhamanan, Saiful Haq,
Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam,
Heather Miller, Matei Zaharia, Christopher Potts

1Stanford University, 2UC Berkeley, 3Carnegie Mellon University,
4Amazon Alexa AI, 5Dashworks Technologies, Inc., 6IIT Bombay,
7Calera Capital, 8Microsoft Qatar, 9Two Sigma Investments

okhattab@cs.stanford.edu

ABSTRACT

The ML community is rapidly exploring techniques for prompting language models (LMs) and for stacking them into pipelines that solve complex tasks. Unfortunately, existing LM pipelines are typically implemented using hard-coded “prompt templates”, i.e. lengthy strings discovered via trial and error. Toward a more systematic approach for developing and optimizing LM pipelines, we introduce DSPy, a programming model that abstracts LM pipelines as text transformation graphs, or imperative computational graphs where LMs are invoked through declarative modules. DSPy modules are parameterized, meaning they can learn how to apply compositions of prompting, finetuning, augmentation, and reasoning techniques. We design a compiler that will optimize any DSPy pipeline to maximize a given metric, by creating and collecting demonstrations. We conduct two case studies, showing that succinct DSPy programs can express and optimize pipelines that reason about math word problems, tackle multi-hop retrieval, answer complex questions, and control agent loops. Within minutes of compiling, DSPy can automatically produce pipelines that outperform out-of-the-box few-shot prompting as well as expert-created demonstrations for GPT-3.5 and Llama2-13b-chat. On top of that, DSPy programs compiled for relatively small LMs like 770M parameter T5 and Llama2-13b-chat are competitive with many approaches that rely on large and proprietary LMs like GPT-3.5 and on expert-written prompt chains. DSPy is available at https://github.com/stanfordnlp/dspy.

1 INTRODUCTION

Language models (LMs) are enabling researchers to build NLP systems at higher levels of abstraction and with lower data requirements than ever before (Bommasani et al., 2021). This is fueling an exploding space of “prompting” techniques—and lightweight finetuning techniques—for adapting LMs to new tasks (Kojima et al., 2022), eliciting systematic reasoning from them (Wei et al., 2022; Wang et al., 2022b), and augmenting them with retrieved sources (Guu et al., 2020; Lazaridou et al., 2022) or with tools (Yao et al., 2022). Most of these techniques are explored in isolation, but interest has been growing in building multi-stage pipelines and agents that decompose complex tasks into more manageable calls to LMs in an effort to improve performance (Qi et al., 2019; Khattab et al., 2021a; Karpas et al., 2022; Dohan et al., 2022; Pourreza & Rafiei, 2023; Yao et al., 2023).

Unfortunately, LMs are known to be sensitive to how they are prompted for each task, and this is exacerbated in pipelines where multiple LM calls have to interact effectively. As a result, the LM calls in existing LM pipelines and in popular developer frameworks are generally implemented using hard-coded ‘prompt templates’, that is, long strings of instructions and demonstrations that are hand crafted through manual trial and error. We argue that this approach, while pervasive, can be brittle
and unscalable—conceptually akin to hand-tuning the weights for a classifier. A given string prompt might not generalize to different pipelines or across different LMs, data domains, or even inputs.

Toward a more systematic approach to designing AI pipelines, we introduce the DSPy programming model. DSPy pushes building new LM pipelines away from manipulating free-form strings and closer to programming (composing modular operators to build text transformation graphs) where a compiler automatically generates optimized LM invocation strategies and prompts from a program. We draw inspiration from the consensus that emerged around neural network abstractions (Bergstra et al., 2013), where (1) many general-purpose layers can be modularly composed in any complex architecture and (2) the model weights can be trained using optimizers instead of being hand-tuned.

To this end, we propose the DSPy programming model (Section 3). We first translate string-based prompting techniques, including complex and task-dependent ones like Chain of Thought (Wei et al., 2022) and ReAct (Yao et al., 2022), into declarative modules that carry natural-language typed signatures. DSPy modules are task-adaptive components—akin to neural network layers—that abstract any particular text transformation, like answering a question or summarizing a paper. We then parameterize each module so that it can learn its desired behavior by iteratively bootstrapping useful demonstrations within the pipeline. Inspired directly by PyTorch abstractions (Paszke et al., 2019), DSPy modules are used via define-by-run computational graphs. Pipelines are expressed by (1) declaring the modules needed and (2) using these modules in any logical control flow (e.g., if statements, for loops, exceptions, etc.) to logically connect the modules.

We then develop the DSPy compiler (Section 4), which optimizes any DSPy program to improve quality or cost. The compiler inputs are the program, a few training inputs with optional labels, and a validation metric. The compiler simulates versions of the program on the inputs and bootstraps example traces of each module for self-improvement, using them to construct effective few-shot prompts or finetuning small LMs for steps of the pipeline. Optimization in DSPy is highly modular: it is conducted by teleprompters, which are general-purpose optimization strategies that determine how the modules should learn from data. In this way, the compiler automatically maps the declarative modules to high-quality compositions of prompting, finetuning, reasoning, and augmentation.

Programming models like DSPy could be assessed along many dimensions; we focus on the role of hand-crafted prompts in shaping system performance. We are seeking to reduce or even remove their role through DSPy modules and teleprompters. We report on two expansive case studies: math word problems (GSM8K; Cobbe et al. 2021) and multi-hop question answering (HotPotQA; Yang et al. 2018) with explorations of chain of thought, multi-chain reflection, multi-hop retrieval, retrieval-augmented question answering, and agent loops. Our evaluations use a number of different compiling strategies effectively and show that straightforward DSPy programs outperform systems using hand-crafted prompts, while also allowing our programs to use much smaller and hence more efficient LMs effectively. Overall, the main contributions of this paper are empirical and algorithmic: with DSPy, we have found that we can implement very short programs that can bootstrap state-of-the-art multi-stage NLP systems using LMs as small as Llama2-13b with no hand-crafted prompts.

2 RELATED WORK

This work is inspired by the role that Torch (Collobert et al., 2002), Theano (Bergstra et al., 2010; 2011; Al-Rfou et al., 2016), Chainer (Tokui et al., 2015), and others played in the development in deep learning by providing powerful abstractions. We are seeking to offer a solid conceptual framework and programming abstractions for what we call foundation model programming. We draw on differentiable programming (Wang et al., 2018) but applied to LM calls rather than neural networks, and borrow syntactic elements from PyTorch (Paszke et al., 2019).

In-context learning (McCann et al. 2018; Radford et al. 2018; Brown et al. 2020) is a key mechanism for programming LMs in this new mode. A growing body of work has revealed that, especially with instruction tuning (Ouyang et al., 2022), we can elicit sophisticated behavior via prompting (Wei et al., 2022; Wang et al., 2022b; Press et al., 2022; Yao et al., 2022; Khot et al., 2022; Madaan et al., 2023). Similarly, forms of weak supervision that would normally require task-specific training are turned into prompts that can bootstrap state-of-the-art multi-stage NLP systems using LMs as small as Llama2-13b with no hand-crafted prompts.

1DSPy is pronounced dee-ess-pie. It’s the second iteration of our earlier Demonstrate–Search–Predict framework (DSP; Khattab et al. 2022).

2We derive the name teleprompters from the notion of abstracting and automating the task of prompting—such it happens at a distance, without manual intervention.
et al., 2021a;b) or hand-built (Ratner et al., 2016; Hancock et al., 2018) heuristics are now done by LMs (Wang et al., 2022b; Zelikman et al., 2022; Zhang et al., 2022; Shao et al., 2023). In-context learning methods now routinely invoke tools, leading to LM pipelines that use retrieval (Chen et al., 2017; Lewis et al., 2020; Guu et al., 2020; Lazaridou et al., 2022; Izacard et al., 2022), and APIs (Nakano et al., 2021). Toolkits have been developed to facilitate this, including LangChain (Chase, 2022), Semantic Kernel (Microsoft, 2023), and Llamalndex (Liu, 2022), and many other retrieval or agent libraries. While these toolkits provide pre-packaged chains and agents that connect LMs with tools, they suffer from the pervasive prompt engineering challenges we address in DSPy: in particular, their abstractions still express task-specific behavior through hand-written prompt templates (see Appendix D). LMQL (Beurer-Kellner et al., 2023) is a related query language, which efficiently constrains the decoding algorithms of LMs to generate only strings that fulfill logical constrains (e.g., lists of bullets or values formatted correctly for a calculator). Whereas DSPy focuses on optimizing LM pipelines for a given metric, LMQL’s lower-level interface for controlled decoding could be beneficial to implement specific advanced modules within DSPy.

Researchers are starting to apply discrete optimization and RL to find effective prompts, generally for a single logical LM call (Guo et al., 2023; Pryzant et al., 2023; Huang et al., 2022; Yang et al., 2023). DSPy seeks to generalize this space: it offers a rich framework for optimizing arbitrary pipelines from high-level declarative signatures, by bootstrapping high-quality multi-stage demonstrations with constraints. In this framework, DSPy teleprompters may apply optimization using model selection techniques like cross-validation or, in principle, with sophisticated techniques involving RL and LM feedback (Hu et al., 2023; Zhao et al., 2023a; Shinn et al., 2023) or learned or Bayesian hyperparameter optimization methods (Bergstra et al., 2013; Akiba et al., 2019). The present paper seeks to motivate DSPy as a programming model and to report new empirical findings from applying the DSPy compiler. This is inspired by formative work by Bergstra et al. (2010; 2013), Paszke et al. (2019), and Wolf et al. (2020), who support their respective programming models with a mix of benchmark numbers and some qualitative measures. For the current paper, we focus on showing that DSPy and its compiler allow us to build outstanding LM systems without hand-crafted prompt strings, but instead from truly modular units, and that this opens up doors for systematically exploring a rich design space at a very high programmatic level of abstraction.

3 The DSPy Programming Model

We present DSPy, which treats LMs as abstract devices for text generation, and optimizes their usage in arbitrary computational graphs. DSPy programs are expressed in Python: each program takes the task input (e.g., a question to answer or a paper to summarize) and returns the output (e.g., an answer or a summary) after a series of steps. DSPy contributes three abstractions toward automatic optimization: signatures, modules, and teleprompters. Signatures abstract the input/output behavior of a module; modules replace existing hand-prompting techniques and can be composed in arbitrary pipelines; and teleprompters optimize each module in the pipeline to maximize a metric.

3.1 Natural Language Signatures Can Abstract Prompting & Finetuning

Instead of free-form string prompts, DSPy programs use signatures to assign work to the LM. This is a natural-language typed function declaration: a declarative spec that tells DSPy what a transformation needs to do (e.g., consume questions and return answers), rather than how a specific LM should be prompted to implement that behavior. More formally, a DSPy signature is a tuple of input fields and output fields (and optional instruction). A field consists of field name and optional metadata. The roles of fields are inferred from field names, e.g. the compiler will use in-context learning to interpret question differently from answer and to iteratively refine its usage of these fields.

Signatures offer two benefits over prompts: they support compilation into high-quality, self-improving, and pipeline-adaptive prompts or finetunes. This is primarily done by bootstrapping (Sec 4) useful demonstrating examples for each signature. Additionally, they handle structured formatting and parsing logic to reduce (or, ideally, avoid) brittle string manipulation in user programs.

---

3We assume access to one or more LMs, which consume a prompt string and return text completions. This may be a promptable LM capable of in-context learning (e.g., GPT-3.5 or Llama2-7b) or a smaller finetunable LM (e.g., T5-base). An LM may be selected as the default; operations will use it unless configured otherwise.

4Description of the task is optional and usually omitted. Fields can carry optional prefix and description.
In practice, DSPy signatures can be expressed with a shorthand notation like `question -> answer`, so that line 1 in the following is a complete DSPy program for a basic question-answering system (with line 2 illustrating usage and line 3 the response when GPT-3.5 is the LM):

```python
1 qa = dspy.Predict("question -> answer")
2 qa(question="Where is Guaraní spoken?") # Out: Prediction(answer='Mainly in South America.')
```

In the shorthand notation, each field’s name indicates the semantic role that the input (or output) field plays in the transformation. DSPy will parse this notation and expand the field names into meaningful instructions for the LM, so that `english_document -> french_translation` would prompt for English to French translation. When needed, DSPy offers more advanced programming interfaces for expressing more explicit constraints on signatures (Appendix B).

### 3.2 Parameterized & Templated Modules can Abstract Prompting Techniques

Akin to type signatures in programming languages, DSPy signatures define an interface and provide type-like hints on the expected behavior. To use a signature, we declare a module with that signature, like the `Predict` module above. A module declaration like this returns a function having that signature. The core module in DSPy is `Predict` (Appendix G.1), which stores internally the supplied signature, an optional LM to use (initially `None`, but otherwise overrides the default LM for this module), and a list of demonstrations for prompting (initially empty). The instantiated module behaves as a callable function: it takes in keyword arguments corresponding to the signature input fields (e.g., `question`), formats a prompt to implement the signature and include the appropriate demonstrations, calls the LM, and parses the output fields. When `Predict` detects it’s being used in compile mode, it internally tracks input/output traces to assist bootstrapping.

DSPy includes more sophisticated modules like `ChainOfThought`, `ProgramOfThought`, `MultiChainComparison`, and `ReAct`. These can be used interchangeably to implement a DSPy signature, e.g. changing `Predict` to `ChainOfThought` above leads to a system that thinks step by step before committing to its output field. Importantly, these modules are implemented in a few lines of code by expanding the user-defined signature and calling `Predict` one or more times on new signatures. We show a simplified implementation of `ChainOfThought` in Appendix G.2 using seven lines of code. This is a fully-fledged module capable of learning effective few-shot prompting for any LM or task. We contrast that with Appendix E, which copies long reasoning prompts hand-written by recent papers and prompting libraries.

DSPy modules translate prompting techniques into modular functions that support any signature by parameterizing these modules. To understand this, observe that any LM call seeking to implement a particular signature needs to specify parameters that include: (1) the specific LM to call (Chen et al., 2023), (2) the prompt instructions (Yang et al., 2023) and the string prefix of each signature field and, most importantly, (3) the demonstrations used as few-shot prompts (for frozen LMs) or as training data (for finetuning). We focus in this paper on generating and selecting useful demonstrations and using them for prompting or finetuning. We find that bootstrapping demonstrations gives us a powerful way to teach sophisticated pipelines of LMs new behaviors systematically.

DSPy modules can be composed in arbitrary pipelines in a define-by-run interface. Inspired directly by PyTorch and Chainer, one first declares the modules needed at initialization, allowing DSPy to keep track of them for optimization, and then one expresses the pipeline with arbitrary code that calls the modules in a forward method. As a simple illustration, we offer the following simple but complete retrieval-augmented generation (RAG) system.

DSPy programs may use tools, which are modules that execute computation. We support retrieval models through a `dspy.Retrieve` module. At the time of writing, DSPy has built-in support for ColBERTv2, Pyserini, and Pinecone retrievers, and we have explored experimental `dspy.SQL` for executing SQL queries and `dspy.PythonInterpreter` for executing Python code in a sandbox.

---

5These generalize prompting techniques from the literature, respectively, by Wei et al. (2022), Chen et al. (2022), Yoran et al. (2023), and Yao et al. (2022).

6To highlight modularity, we use `ChainOfThought` as a drop-in replacement of the basic `Predict`. One can now simply write `RAG()("Where is Guaraní spoken?")` to use it. Notice that, if we use a signature "context, question -> search_query", we get a system that generates search queries rather than answers.
class RAG(dspy.Module):
    def __init__(self, num_passages=3):
        self.retrieve = dspy.Retrieve(k=num_passages)
        self.generate_answer = dspy.ChainOfThought("context, question -> answer")
    def forward(self, question):
        context = self.retrieve(question).passages
        return self.generate_answer(context=context, question=question)

3.3 Teleprompters Can Automate Prompting for Arbitrary Pipelines

When compiling a DSPy program, we generally invoke a teleprompter, which is an optimizer that takes the program, a training set, and a metric—and returns a new optimized program. Different teleprompters (Sec 4) apply different strategies for optimization, typically gradient-free. In DSPy, training sets may be small, potentially a handful of examples, though larger data enables more powerful optimization. Training examples may be incomplete, i.e., only input values are necessary. Labels for the pipeline steps are not required, unless they need to be used in the metric. In practice, we typically assume labels only for (at most) the program’s final output, not the intermediate steps. This label-efficiency is critical for modularity: building a new pipeline in DSPy requires simply recompiling the new pipeline’s code, not annotating data specific to the new pipeline.

Metrics can be simple notions like exact match (EM) or F1, or they can be entire DSPy programs that balance multiple concerns. For example, we may compile the RAG module above against a dataset of question–answer pairs qa_trainset and the metric EM. The goal of optimization here is to effectively bootstrap few-shot demonstrations. The following code achieves this:

```python
# Small training set with only questions and final answers.
qa_trainset = [dspy.Example(question="What is the capital of France?", answer="Paris")]

# The teleprompter will bootstrap missing labels: reasoning chains and retrieval contexts.
compiled_rag = dspy.BootstrapFewShot(metric=dspy.evaluate.answer_exact_match).compile(RAG(), trainset=qa_trainset)
```

In this example, the BootstrapFewShot teleprompter (Sec 4, Appendix G.2) simulates RAG on the training example(s). It will collect demonstrations of each module (i.e., examples of its input–output behavior) that collectively lead to valid output (i.e., respecting the signatures and the metric).

Teleprompters can be composed by specifying a teacher program. DSPy will sample demonstrations from this program for prompt optimization. This composition can enable very rich pipelines, where expensive programs (e.g., ensembles of large LMs) supervise cheap programs (e.g., simpler pipelines using smaller LMs). One may start with compiled_rag from above (say, compiled to use a large Llama2-13b-chat LM) but now fine-tune Flan-T5-large to create an efficient program:

```python
# Larger set of questions with no labels*. Labels for all steps will be bootstrapped.
unlabeled_questions = [dspy.Example(question="What is the capital of Germany?")]

# As we assume no answer, we use 'answer_passage_match' to filter ungrounded answers.
compiled_rag_via_finetune = dspy.FinetuneTeleprompter.compile(RAG(), teacher=compiled_rag, trainset=unlabeled_questions, target="google/flan-t5-large")
```

4 The DSPy Compiler

A key source of DSPy’s expressive power is its ability to compile—or automatically optimize—any program. Compiling relies on a teleprompter, which is an optimizer for DSPy programs that improves the quality (or cost) of modules via prompting or finetuning, which are unified in DSPy. The compiler first finds all unique Predict modules (predictors) in a program, including those nested under other modules. While DSPy does not enforce this when creating new teleprompters, typical teleprompters go through three stages.

Stage 1: Candidate Generation For each predictor p, the teleprompter may generate candidate values for the parameters of p: the instructions, field descriptions, or demonstrations (i.e., example input–output pairs). In this iteration of DSPy, we focus on demonstrations and find that simple rejection sampling can help bootstrap highly effective multi-stage systems. Consider the simplest non-trivial teleprompter in DSPy, BootstrapFewShot (simplified pseudocode in Appendix H.1).
This teleprompter will simulate a teacher program (or, if unset, the zero-shot version of the program) on some training inputs, possibly one or more times with a high temperature. When running in compile mode, multi-stage traces are tracked transparently and in a thread-safe fashion throughout execution. The program’s metric is used to filter for multi-stage traces that together help the pipeline pass the metric. We thus obtain potential labels for all signatures in the program by throwing away the bad examples and using the good examples as potential demonstrations, though these design decisions are under user control. While LMs can be highly unreliable, we find they can be rather efficient at searching the space of solutions for multi-stage designs. A well-decomposed program can typically find at least a few training examples where the LM can pass the constraints enforced by the signatures and metrics, allowing us to bootstrap iteratively if needed.

Stage 2: Parameter Optimization

Now each parameter has a discrete set of candidates: demonstrations, instructions, etc. Many hyperparameter tuning algorithms (e.g., random search or Tree-structured Parzen Estimators as in HyperOpt (Bergstra et al., 2013) and Optuna (Akiba et al., 2019)) can be applied for selection among candidates. We report a simplified implementation of DSPy’s BootstrapFewShotWithRandomSearch in Appendix H.2. Another type of optimization is finetuning with BootstrapFinetune, where the demonstrations are used to update the LM’s weights for each predictor. When this is applied, the LM of each module is updated to the new LM weights.

Stage 3: Higher-Order Program Optimization

A different type of optimization that the DSPy compiler supports is modifying the control flow of the program. One of the simplest forms of this is ensembling, which we use in the case studies in this work. An ensemble will bootstrap multiple copies of the same program, and then replace the program with a new one that runs them all in parallel and reduces their predictions into one with a custom function (e.g., majority voting). In future work, this stage can easily accommodate techniques for more dynamic (i.e., test-time) bootstrapping as well as automatic backtracking-like logic.

5 Goals of Evaluation

Programming frameworks can be evaluated along many dimensions: computational efficiency, developer efficiency, intuitiveness of the code and concepts, and so forth. In this paper, we focus on perhaps the most pressing issue for current LM pipelines: the role of hand-written, task-specific prompts in achieving performant systems. Our evaluations seek to test the following hypotheses:

H1 With DSPy, we can replace hand-crafted prompt strings with concise and well-defined modules, without reducing quality or expressive power.

H2 Parameterizing the modules and treating prompting as an optimization problem makes DSPy better at adapting to different LMs, and it may outperform expert-written prompts.

H3 The resulting modularity makes it possible to more thoroughly explore complex pipelines that have useful performance characteristics or that fit nuanced metrics.

Our evaluation will explore these hypotheses using diverse task–program pairs. We hope this begins a shift from underspecified questions like “how do different LMs compare on GSM8K” toward “how they compare on GSM8K with program P when compiled with strategy S”, which is a well-defined and reproducible run. Ultimately, our goal is to reduce the role of artful prompt construction in modern AI, in favor of the development of new modular, composable programs and optimizers.

6 Case Study: Math Word Problems

We evaluate on the popular GSM8K dataset with grade school math questions (Cobbe et al., 2021). We sample 200 and 300 question–answer pairs from the official training set for training and development, respectively. Our final evaluations use the 1.3k official test set examples. We report extensive comparisons on the development set to avoid overfitting on test. Following prior work on GSM8K, we evaluate the accuracy of the final numerical value that appears in the LM output.

Programs Considered

For this task, we consider three simple DSPy programs: a one-step Predict module (vanilla), a two-step ChainOfThought module (CoT), and finally a multi-stage ComparatorOfThoughts module (ThoughtReflection). These are defined by the code below. In reflection, five reasoning chains are sampled from the LM (alongside their answers) and they are compared in parallel by a built-in MultiChainComparison module, which generalizes Yoran.
Table 1: Results with in-context learning on GSM8K math word problems. Each row represents a separate pipeline: the module in the Program column is compiled against the examples in the Training set. The programs, compilers, and (small) training sets are defined in Section 6. Rows with ensemble build on the immediately preceding row. Notably, all programs in this table are expressed by composing two to four DSPy modules and teleprompters. Compiling the right modules, instead of string prompts, improves different LMs from 9–25% accuracy to 46–81% accuracy.

<table>
<thead>
<tr>
<th>Program</th>
<th>Compilation</th>
<th>Training</th>
<th>GPT-3.5 Dev</th>
<th>GPT-3.5 Test</th>
<th>Llama2-13b-chat Dev</th>
<th>Llama2-13b-chat Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>vanilla</td>
<td>none</td>
<td>n/a</td>
<td>24.0</td>
<td>25.2</td>
<td>7.0</td>
<td>9.4</td>
</tr>
<tr>
<td>vanilla</td>
<td>fewshot</td>
<td>trainset</td>
<td>33.1</td>
<td>4.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vanilla</td>
<td>bootstrap</td>
<td>trainset</td>
<td>44.0</td>
<td>28.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vanilla</td>
<td>+ensemble</td>
<td>trainset</td>
<td>64.7</td>
<td>61.7</td>
<td>37.3</td>
<td>36.5</td>
</tr>
<tr>
<td>CoT</td>
<td>none</td>
<td>n/a</td>
<td>56.0</td>
<td>26.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoT</td>
<td>fewshot</td>
<td>trainset</td>
<td>65.1</td>
<td>27.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoT</td>
<td>bootstrap</td>
<td>trainset</td>
<td>78.6</td>
<td>72.4</td>
<td>34.3</td>
<td>33.7</td>
</tr>
<tr>
<td>CoT</td>
<td>+ensemble</td>
<td>trainset</td>
<td>80.3</td>
<td>72.9</td>
<td>43.3</td>
<td></td>
</tr>
<tr>
<td>reflection</td>
<td>none</td>
<td>n/a</td>
<td>65.0</td>
<td>36.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reflection</td>
<td>fewshot</td>
<td>trainset</td>
<td>71.7</td>
<td>36.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reflection</td>
<td>bootstrap</td>
<td>trainset</td>
<td>83.0</td>
<td>76.0</td>
<td>44.3</td>
<td>40.2</td>
</tr>
<tr>
<td>reflection</td>
<td>+ensemble</td>
<td>trainset</td>
<td>86.7</td>
<td>49.0</td>
<td>49.0</td>
<td>46.9</td>
</tr>
</tbody>
</table>

et al. (2023). This generates a new answer taking into account the patterns from the five attempts. Critically, the modules used are all generic, none is specific to math problems or a particular LM.

```python
1 vanilla = dspy.Predict("question -> answer") # GSM8K Program 'vanilla'
2 CoT = dspy.ChainOfThought("question -> answer") # GSM8K Program 'CoT'
3
class ThoughtReflection(dspy.Module):
4 def __init__(self, num_attempts):
5     self.predict = dspy.ChainOfThought("question -> answer", n=num_attempts)
6     self.compare = dspy.MultiChainComparison("question -> answer", M=num_attempts)
7
def forward(self, question):
8     completions = self.predict(question=question).completions
9     return self.compare(question=question, completions=completions)
10
11 reflection = ThoughtReflection(num_attempts=5) # GSM8K Program 'reflection'
```

**Compiling** As we discussed in Section 4, DSPy programs can be compiled into new, optimized programs. In our experiments, we evaluate the programs zero-shot (no compiling) as well as a number of strategies for compiling. Our simplest compiler is LabeledFewShot:

```python
1 fewshot = dspy.LabeledFewShot(k=8).compile(program, trainset=trainset)
```

Here, program can be any DSPy module. This samples k=8 demonstrations from the trainset for the fields common to the training examples and the signature(s), in this case, question and answer but not reasoning for instance. We report the average of 3–5 runs (depending on the setting) when applying such sampling. Next, we consider bootstrapping few-shot examples with random search:

```python
1 tp = BootstrapFewShotWithRandomSearch(metric=gsm8k_accuracy)
2 bootstrap = tp.compile(program, trainset=trainset, valset=devset)
```

This will generate demonstration chains for examples in the training set and optimize the selection of demonstrations (from this set) to self-improve the program’s modules. As the name indicates, this is done with random search, treating the selection of demonstrations as a parameter to optimize. Next, if desired, this bootstrapping process can be nested in DSPy. In particular, we can use the optimized bootstrap program itself to further bootstrap another program. This is relevant, for example, whenever the original zero-shot program performs relatively poorly.

```python
1 bootstrap2 = tp.compile(program, teacher=bootstrap, trainset=trainset, valset=devset)
```

And lastly, we consider ensembling these bootstraps:

```python
1 # A program that ensembles the top-7 candidate programs from a bootstrapping compiler run
2 # (in particular 'bootstrap' or, when applicable, 'bootstrap2') with majority voting.
3 ensemble = Ensemble(reduce_fn=dspy.majority).compile(bootstrap.programs[:7])
```
GSM8K includes human reasoning chains. Above, trainset does not include these reasoning chains. We also evaluate with trainset_human_CoT, which extends the examples in trainset with the human reasoning string. These two datasets can be used interchangeably as the value for the trainset parameter above. We note here that compiling generally runs on the order of minutes (or tens of minutes) as even the more expensive settings only require running the program a few thousand times (e.g., 10–20 trials over 150–300 validation examples) and they can occur in parallel.

Results Our results are summarized in Table 1, which includes dev results as well as our evaluation of promising representatives of each approach on the test set. First, the reflection program while only a few lines longer than the others is a clear winner, though CoT is quite effective with ensemble. Second, the bootstrap compilation procedure leads to large gains for every program, across both LMs. Interestingly, vanilla is helped by compiling with bootstrap as the teacher program (bootstrap×2). On inspecting the prompts bootstrapped (Appendix I), we see that the LM leverages the metric (i.e., correctness of the final numerical value in the output) so it uses the answer field for reasoning first. Third, while the human reasoning chains (+human_CoT) provide a large boost for fewshot, we can match or surpass this using bootstrap, which substantiates our hypothesis that DSPy can cut the need for hand-crafted prompts.

7 CASE STUDY: COMPLEX QUESTION ANSWERING

In this case study, we explore the multi-hop question answering task with the HotPotQA (Yang et al., 2018) dataset in the open-domain “fullwiki” setting. For retrieval, we use a search index of the official Wikipedia 2017 “abstracts” dump of HotPotQA. Search is conducted by a ColBERTv2 (Santhanam et al., 2021) retriever. The HotPotQA test set is hidden, so we reserve the official validation set for our testing, and sample 1000 examples for that. We sub-divide the training set into 70%/30% train/validation splits. In the training (and thus validation) split, we keep only examples marked as “hard” in the original dataset, which matches the designation of the official validation and test sets. For training and for reporting development results, we sample 200 and 300 examples respectively.

Programs Considered Our simplest baseline is the vanilla program used in the previous case study on GSM8K (Section 6); the "question -> answer" signature is universal enough that it will work for many tasks when compiled appropriately. Our baseline RAG program is the one given in Section 3.2 with a dsyp.ChainOfThought layer. This program does not excel at HotPotQA, and this motivates us to evaluate two multi-hop programs. We first test ReAct (Yao et al., 2022), a multi-step agent for tool use, which is implemented as a built-in module in DSPy. In the simplest case, a ReAct module for a particular signature can be declared as follows in DSPy:

```python
react = dsyp.ReAct("question -> answer", tools=[dsyp.Retrieve(k=1)], max_steps=5)
```

And we test the following custom program, which is akin to Baleen (Khattab et al., 2021a), IRRR (Qi et al., 2020), and has similarities to IRCoT (Trivedi et al., 2022):

```python
class BasicMultiHop(dspy.Module):
    def __init__(self, passages_per_hop):
        self.retrieve = dsyp.Retrieve(k=passages_per_hop)
        self.generate_query = dsyp.ChainOfThought("context, question -> search_query")
        self.generate_answer = dsyp.ChainOfThought("context, question -> answer")

    def forward(self, question):
        context = []
        for hop in range(2):
            query = self.generate_query(context=context, question=question).search_query
            context += self.retrieve(query).passages
        return self.generate_answer(context=context, question=question)

multihop = BasicMultiHop(passages_per_hop=3)
```

7 We can informally compare these with published results for GSM8K. Zhang et al. (2022) reports 48% for text-davinci-002, close to our results using llama2-13b, and 63% with codex using CoT. Zhang et al. (2022) report 57% for CoT prompting with PaLM 540-B, which becomes 74% upon adding self-consistency. Llama2 (Touvron et al., 2023) presented 28.7%, 42.2%, and 56.8% for llama2-13b, llama2-34b, and llama2-70b. Our implementation with llama2-13b is competitive with their llama2-34b results, while we don’t use human reasoning chains in our program. Zhao et al. (2023b) reports that CoT scores around 80% for gpt-3.5-turbo from April 2023. The GPT-4 paper (OpenAI, 2023) reports that GPT-3.5 scores 57% and GPT-4 elevates this to 92% but they note that GPT-4 was trained on a subset of GSM8K.
Table 2: Results with in-context learning on HotPotQA. We report answer exact match (Ans) and pair-retrieval accuracy (Psg). Each row represents a separate pipeline: the module in the Program column is compiled against the examples in the Training set. The programs, compilers, and (small) training sets are defined in the main text. For HotPotQA, we use the training set (and not dev) directly for cross-validation. **The marked result is evaluated on 50% of our test set due to cost.

<table>
<thead>
<tr>
<th>Program</th>
<th>Compiler</th>
<th>GPT-3.5</th>
<th>Llama2-13b-chat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dev Ans</td>
<td>Test Ans</td>
</tr>
<tr>
<td>vanilla</td>
<td>fewshot</td>
<td>34.3 n/a</td>
<td>31.5 n/a</td>
</tr>
<tr>
<td>CoT_RAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fewshot</td>
<td>36.4 36.0</td>
<td>29.8 34.4</td>
</tr>
<tr>
<td></td>
<td>bootstrap</td>
<td>42.3 36.0</td>
<td>– –</td>
</tr>
<tr>
<td>react</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>20.3 – –</td>
<td>– –</td>
</tr>
<tr>
<td></td>
<td>+human_r</td>
<td>33.0 – –</td>
<td>– –</td>
</tr>
<tr>
<td></td>
<td>bootstrap</td>
<td>31.0 – –</td>
<td>– –</td>
</tr>
<tr>
<td></td>
<td>bootstrapx2</td>
<td>39.0 – –</td>
<td>– –</td>
</tr>
<tr>
<td>multihop</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>fewshot</td>
<td>36.9 38.3</td>
<td>31.2 40.8</td>
</tr>
<tr>
<td></td>
<td>bootstrap</td>
<td>48.7 47.0</td>
<td>39.6 43.8</td>
</tr>
<tr>
<td></td>
<td>ensemble</td>
<td>54.7 –</td>
<td>45.6* –</td>
</tr>
</tbody>
</table>

**Compiling** We continue to use the compilers we used for GSM8K (see Section 6). We also consider two compositions of our teleprompters. For ReAct, we consider applying BootstrapFewShotWithRandomSearch starting from an earlier bootstrap of the ReAct program. For the simple multihop program, we consider fine-tuning starting from its earlier bootstrap.

```python
multihop_t5 = dspy.BootstrapFinetune(metric=answer_exact_match).compile(program,
teacher=bootstrap, trainset=trainset, target='t5-large')
```

**Results** Table 2 summarizes our results: a simple multihop program is the best across all models and in general bootstrap again proves to be very effective. Perhaps most importantly, we can make Llama2-13b-chat competitive with GPT-3.5 using the multihop program. We also evaluated the compiler multihop_t5 defined above which produces a T5-Large (770M parameter) model. This program scores 39.3% answer EM and 46% passage accuracy on the dev set, using only 200 labeled inputs and 800 unlabeled questions otherwise, with a teacher program consisting of an ensemble (union) of two Llama2-13b-chat multihop programs. Such a program would impose orders of magnitude lower costs for inference than a proprietary LM like GPT-3.5.6

8 Conclusion

This paper introduced DSPy, a new programming model for designing AI systems using pipelines of pretrained LMs and other tools. We presented three new concepts introduced in this abstraction (DSPy signatures, modules, and teleprompters), and showed in two very different case studies that it supports rapid development of highly effective systems that use relatively small LMs. We have maintained open-source versions of this framework throughout 2023 and since. In this period, we have seen and created a large number of programs that were compiled to high-quality systems by DSPy, spanning tasks from information extraction to low-resource synthetic data generation. In the interest of space and to maintain reasonable scope in this paper, we leave reporting on such tasks under controlled experimental conditions to future work. While in-context learning has proved transformative, we argue that the true expressive power in this emerging paradigm is in building sophisticated text transformation graphs in which composable modules and optimizers (teleprompters) come together to leverage LMs in more systematic and reliable ways.

---

6Our results may be pegged against the evaluation on HotPotQA in a number of recent papers, though there can be significant variation in evaluation methodology and test set samples across studies in this space. Si et al. (2022) achieve 25.2% EM with CoT prompting. With a “recite-and-answer” technique for PaLM-62B (Chowdhery et al., 2022), Sun et al. (2022) achieve 26.5% EM. Wang et al. (2022a) achieve 33.8% EM and 44.6 F1 when applying self-consistency for PaLM-540B. Yao et al. (2022) achieve 35.1% EM using the ReAct agent, with a tool giving it the ability for search using a Wikipedia API. Trivedi et al. (2022) reports 49% using a pipeline with code-davinci-002 on a sample of 500 HotPotQA questions.
REFERENCES


Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Efficiently programming large language models using sglang, 2023.
A  Latency & Throughput

In the main paper, we introduced the DSPy programming model and presented quality-oriented case studies. In this appendix, we report on efficiency-oriented metrics. We select one of the most sophisticated programs we discussed, namely, the multihop program from Sec 7.

We run a self-contained experiment from scratch to sidestep the effect of caching. We compile the multihop program using the latest OpenAI GPT-3.5 model at the time of writing this appendix, i.e. gpt-3.5-turbo-1106. We compile using the BootstrapFewShotWithRandomSearch teleprompter with 7 candidate programs. DSPy supports parallel evaluation. In these runs, we set the maximum number of parallel threads to 10.

Compiling this program takes approximately 370 seconds (or 6 minutes), which is a little shorter than one minute per candidate program. The number of candidates for random search is a simple configurable parameter, whose typical range is 5 through 30. For OpenAI models, our main experiments use 13 candidate, though we observe strong results with as little as 7. This compilation process makes approximately 3200 API calls to the language model, with a total of 2.7M input tokens and 156k output tokens. With the current pricing of OpenAI models, this costs a total of $3.0 USD to compile.

For inference, single-threaded inference with the multihop program takes on the order of 2–3 seconds per question, which includes three chain of thought prompts (i.e., two for query generation and one for answer generation) and two retrieval queries (i.e., each retrieving three passages). In the presence of multiple questions, multi-threading is again possible, and we observe that we can process more than 150 questions per minute with 10 threads. The exact maximum depends on the rate limits supported by the retriever and the LM server.

To provide a comparison of inference costs, we took four rows of increased complexity for HotPotQA (Table 2) and tested them with gpt-3.5-1106, while disabling the cache (which would otherwise skip any repeated computations, e.g. retrieval queries or LM calls). We ran 100 questions from HotPotQA with a single thread, and we report the average latency below. Each numbered bullet below represents a program (with compiler strategy in parentheses). For Vanilla (fewshot), we observe average latency of 0.3 seconds per question and average Cost of $0.0005 per question (1x). For CoT_RAG (fewshot), we observe average latency of 1.1 seconds per question and average cost of $0.0013 per question (2.6x). For Multihop (fewshot), we observe average latency of 2.6 seconds per question and average cost of $0.0018 per question (3.6x). For Multihop (bootstrap), we observe average latency of 2.6 seconds per question and average cost of $0.0041 per question (8.2x). As this shows, these programs are within an order of magnitude of the cost and latency of the simplest one, even though we see major quality improvements from vanilla to multihop.

Latency and throughput results for gpt-3.5-turbo-1106 depend on the OpenAI API load and the rate limit permitted. For local models at the scale of 13B parameters, we typically see higher latencies and lower throughputs in practice, when serving the model from one or a few A100 GPUs. We suspect this gap will be bridged with recent open research advancements in serving open models, especially approaches that apply intelligent prefix caching for prompts that share the same prefixes, e.g. Zheng et al. (2023).

B  Advanced Signatures

When more control is desired, one can express signatures as Python classes to provide explicit instructions of the transformation and describe the format or role of each field more directly. For instance, the following signature generates search queries using context and an optional question:

```python
class GenerateSearchQuery(dspy.Signature):
    """Write a simple search query that will help answer a complex question."""
    context = dspy.InputField(desc="may contain relevant facts")
    question = dspy.InputField()
    query = dspy.OutputField(prefix="Search Query:")
```

Using the above, we can specify a complete system for the generation of a synthetic IR dataset where the queries are mediated by a question generated by the LM:
query_gen = dspy.Predict(GenerateSearchQuery)
query_gen(context="Language typology")
# Out: Prediction(question='What are the main types of language classification?', query="'language classification' OR 'language typology' -wikipedia")

If questions are available, they can be supplied as shown: query_gen(context="Language typology", question="What are the primary language families of South America?").

C More Advanced Metrics

If one wanted to push the compiled program to be extractive given its retrieved contexts, one could define a custom metric to use in place of dspy.evaluate.answer_exact_match:

```python
def my_rag_validation_logic(example, pred, trace=None):
    answer_match = dspy.evaluate.answer_exact_match(example, pred)
    context_match = any((pred.answer.lower() in c) for c in pred.context)
    return answer_match and context_match
```

D Comparison with existing libraries like LangChain and LlamaIndex

LangChain and LlamaIndex are perhaps the most popular library in the general space of prompting LMs. These libraries have a different focus compared to DSPy and they suffer internally from the prompt engineering challenges that DSPy aims to resolve. In particular, whereas the goal of DSPy is to tackle the fundamental challenges of prompt engineering for building new LM computational graphs, LangChain and LlamaIndex generally help application developers who need pre-packaged components and chains, e.g., implementations of popular and reusable pipelines (e.g., particular agents and specific retrieval pipelines) and tools (e.g., connections to various databases and implementations of long- and short-term memory for agents).

These off-the-shelf higher-level abstractions contrast with DSPy’s focus on introducing core composable operators. In particular, DSPy introduces signatures (to abstract prompts), modules (to abstract prompting techniques), and teleprompters to act as optimizers for arbitrary imperative code (DSPy programs) that chain modules together. Its goal is to help researchers and practitioners build new LM pipelines quickly and achieve very high quality through automatic compilation (self-improvement) instead of manual prompt engineering.

In contrast, typical existing research implementations and existing libraries like LangChain and LlamaIndex are implemented using manual prompt engineering, which is the key problem that DSPy tackles. We conducted an informal study to highlight this. In late September 2023, we found that the LangChain codebase contains 50 strings exceeding 1000 characters, which are generally prompts, compared to none at all in DSPy. Indeed, a substantial number of LangChain’s Python files are singularly dedicated to task-related templating and prompt engineering with 12 prompts.py files and and 42 prompt.py files. DSPy, on the other hand, provides a structured framework that automatically bootstraps prompts. The library itself does not contain a single hand-written prompt demonstration for any tasks at the time of writing, despite the very high quality with various LMs.

To review the typical forms of prompt engineering in existing libraries, we consider the following in LangChain. The LangChain Program-Aided Language Model Gao et al. (2023a) chain program uses few-shot learning, leveraging a template that is 3982 characters long with 8 math word problems (Prompt 2) and corresponding outputted programs as learning examples for the language model. LangChain also contains a prompt for SQL query tasks for each of the databases like Oracle, GoogleSQL, DuckDB, Crate, and MySQL, with the average length of these prompts at 1058 characters. Other task areas such as QA with sources (Prompt D) and GraphQA also have significantly lengthy prompt templates, with averages of 1337 and 722 characters, respectively. While expert-written prompts can be useful, we believe that LM- and task-adequate prompts bootstrapped automatically can offer far more power (and are far more modular) than hard-coding a prompt per database provider inside the code base. The next appendix section contains a number of prompts copied from related research papers and existing libraries.
### Sample Large Prompts

This section highlights a few popular existing frameworks that structure prompts with extensive prompt engineering templates. The primary objective is to capture how many words and characters are used for such large multi-line prompts defined for tasks or tools and present these example prompts retrieved from open-sourced papers and repositories. The formatting of these example prompts is adapted from Gao et al. (2023a).

<table>
<thead>
<tr>
<th>Task/Tool Prompt</th>
<th>Source</th>
<th>Words</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt 1: Text-evidence checker</td>
<td>Gao et al. (2023a)</td>
<td>818</td>
<td>4964</td>
</tr>
<tr>
<td>Prompt 2: Math word problems (PAL)</td>
<td>LangChain &amp; Gao et al. (2023b)</td>
<td>566</td>
<td>3957</td>
</tr>
<tr>
<td>Prompt 3: ReAct</td>
<td>Yao et al. (2022)</td>
<td>593</td>
<td>3889</td>
</tr>
<tr>
<td>Prompt 4: Zero-shot ReAct</td>
<td>LangChain</td>
<td>101</td>
<td>600</td>
</tr>
<tr>
<td>Prompt 5: QA with sources</td>
<td>LangChain</td>
<td>992</td>
<td>6197</td>
</tr>
<tr>
<td>Prompt 6: SQL MyScale querying</td>
<td>LangChain</td>
<td>343</td>
<td>2239</td>
</tr>
<tr>
<td>Prompt 7: Relevant docs retrieval</td>
<td>LlamaIndex</td>
<td>129</td>
<td>719</td>
</tr>
<tr>
<td>Prompt 8: IRS chatbot</td>
<td>LlamaIndex</td>
<td>389</td>
<td>2258</td>
</tr>
</tbody>
</table>
Figure 1: Example few-shot prompt using a reasoning chain for agreement model that identifies inconsistencies between text and evidence (Gao et al., 2023a).
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?

```python
def solution():
    # "Olivia has $23. She bought five bagels for $3 each. How much money does she have left?"
    money_initial = 23
    bagels = 5
    bagel_cost = 3
    money_spent = bagels * bagel_cost
    money_left = money_initial - money_spent
    result = money_left
    return result
```

Q: Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?

```python
def solution():
    # "Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?"
    golf_balls_initial = 58
    golf_balls_lost_tuesday = 23
    golf_balls_lost_wednesday = 2
    golf_balls_left = golf_balls_initial - golf_balls_lost_tuesday - golf_balls_lost_wednesday
    result = golf_balls_left
    return result
```

Q: There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?

```python
def solution():
    # "There were nine computers in the server room. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?"
    computers_initial = 9
    computers_per_day = 5
    num_days = 4
    computers_added = computers_per_day * num_days
    computers_total = computers_initial + computers_added
    result = computers_total
    return result
```

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

```python
def solution():
    # "Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?"
    toys_initial = 5
    mom_toys = 2
    dad_toys = 2
    total_received = mom_toys + dad_toys
    total_toys = toys_initial + total_received
    result = total_toys
    return result
```

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

```python
def solution():
    # "Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?"
```
def solution():
    '''Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?'''
    jason_lollipops_initial = 20
    jason_lollipops_after = 12
    denny_lollipops = jason_lollipops_initial - jason_lollipops_after
    result = denny_lollipops
    return result

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

# solution in Python:

def solution():
    '''Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?''
    leah_chocolates = 32
    sister_chocolates = 42
    total_chocolates = leah_chocolates + sister_chocolates
    chocolates_eaten = 35
    chocolates_left = total_chocolates - chocolates_eaten
    result = chocolates_left
    return result

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

# solution in Python:

def solution():
    '''If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?''
    cars_initial = 3
    cars_arrived = 2
    total_cars = cars_initial + cars_arrived
    result = total_cars
    return result

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

# solution in Python:

def solution():
    '''There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?''
    trees_initial = 15
    trees_after = 21
    trees_added = trees_after - trees_initial
    result = trees_added
    return result

Q: {question}

# solution in Python:
Solve a question answering task with interleaving Thought, Action, Observation steps. Thought can reason about the current situation, and Action can be three types:

1. (1) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.
2. (2) Lookup[keyword], which returns the next sentence containing keyword in the current passage.
3. (3) Finish[answer], which returns the answer and finishes the task.

Here are some examples.

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas.
Action 2: Lookup[eastern sector]
Observation 2: (Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains orogeny.
Action 3: Search[High Plains]
Observation 3: High Plains refers to one of two distinct land regions:
Action 4: Search[High Plains (United States)]
Observation 4: The High Plains are a subregion of the Great Plains. From east to west, the High Plains rise in elevation from around 1,800 to 7,000 ft (550 to 2,130 m). [3]
Action 5: Finish[1,800 to 7,000 ft]

Question: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Action 1: Search[Milhouse]
Observation 1: Milhouse Mussolini Van Houten is a recurring character in the Fox animated television series The Simpsons voiced by Pamela Hayden and created by Matt Groening.
Action 2: Lookup[named after]
Observation 2: (Result 1 / 1) Milhouse was named after U.S. president Richard Nixon, whose middle name was Milhous.
Action 3: Finish[Richard Nixon]

Question: Which documentary is about Finnish rock groups, Adam Clayton Powell or The Saimaa Gesture?
Action 1: Search[Adam Clayton Powell]
Action 2: Search[Adam Clayton Powell (film)]
Observation 2: Adam Clayton Powell is a 1989 American documentary film directed by Richard Kilberg. The film is about the rise and fall of influential African-American politician Adam Clayton Powell Jr. [3][4] It was later aired as part of the PBS series The American Experience.
Action 3: Finish[The Saimaa Gesture]

Question: What profession does Nicholas Ray and Elia Kazan have in common?
Action 1: Search[Nicholas Ray]
Observation 1: Nicholas Ray (born Raymond Nicholas Kienzle Jr., August 7, 1911 { June 16, 1979) was an American film director, screenwriter, and actor best known for the 1955 film Rebel Without a Cause.
Action 2: Search[Elia Kazan]
Observation 2: Elia Kazan was an American film and theatre director, producer, screenwriter and actor.
Action 3: Finish[director, screenwriter, actor]

Question: Which magazine was started first, Arthur’s Magazine or First for Women?
Action 1: Search[Arthur’s Magazine]
Observation 1: Arthur’s Magazine (1844–1846) was an American literary periodical published in Philadelphia in the 19th century.
Action 2: Search[First for Women]
Observation 2: First for Women is a women’s magazine published by Bauer Media Group in the USA.[1] The magazine was started in 1989.
Action 3: Finish[Arthur’s Magazine]

Question: Were Pavel Urysohn and Leonid Levin known for the same type of work?
Action 1: Search[Pavel Urysohn]
Observation 1: Pavel Samuilovich Urysohn (February 3, 1898 & August 17, 1924) was a Soviet mathematician who is best known for his contributions in dimension theory.
Action 2: Search[Leonid Levin]
Observation 2: Leonid Anatolievich Levin is a Soviet-American mathematician and computer scientist.
Action 3: Finish[yes]

Figure 3: ReAct example prompt for interleaving Thought, Action, Observation steps.
Figure 4: Langchain ReAct example prompt for interleaving Thought, Action, Observation steps.
Given the following extracted parts of a long document and a question, create a final answer with references ("SOURCES"). If you don’t know the answer, just say that you don’t know. Don’t try to make up an answer.

ALWAYS return a "SOURCES" part in your answer.

QUESTION: Which state/country’s law governs the interpretation of the contract?

Content: This Agreement is governed by English law and the parties submit to the exclusive jurisdiction of the English courts in relation to any dispute (contractual or non-contractual) concerning this Agreement save that either party may apply to any court for an injunction or other relief to protect its Intellectual Property Rights.

Source: 28-pl

11.7 Severability. The invalidity, illegality or unenforceability of any term (or part of a term) of this Agreement shall not affect the continuation in force of the remainder of the term (if any) and this Agreement.

Source: 30-pl

Content: (b) if Google believes, in good faith, that the Distributor has violated or caused Google to violate any Anti-Bribery Laws (as defined in Clause 8.5) or that such a violation is reasonably likely to occur,

Source: 4-pl

FINAL ANSWER: This Agreement is governed by English law.

SOURCES: 28-pl

20 QUESTION: What did the president say about Michael Jackson?

Content: Madam Speaker, Madam Vice President, our First Lady and Second Gentleman, Members of Congress and the Cabinet. Justices of the Supreme Court. My fellow Americans.

Last year COVID-19 kept us apart. This year we are finally together again. Tonight, we meet as Democrats Republicans and Independents. But most importantly as Americans.

Together, let’s imagine a better tomorrow for all Americans. Tonight, I can announce that the United States has worked with 30 other countries to release 60 Million barrels of oil from reserves around the world.

OFFICIAL ANSWER: Madam Speaker, Madam Vice President, our First Lady and Second Gentleman, Members of Congress and the Cabinet. Justices of the Supreme Court. My fellow Americans.

Last year COVID-19 kept us apart. This year we are finally together again. Tonight, we meet as Democrats Republicans and Independents. But most importantly as Americans.

Together, let’s imagine a better tomorrow for all Americans.

Tonight, I can announce that the United States has worked with 30 other countries to release 60 Million barrels of oil from reserves around the world.

Content: And a proud Ukrainian people, who have known 30 years of independence, have repeatedly shown that they will not tolerate anyone who tries to take their country backwards.

Source: 0-pl

Content: And we won’t stop.

Source: 8-pl

Content: No Waiver. Failure or delay in exercising any right or remedy under this Agreement shall not constitute a waiver of such (or any other) right or remedy.

Source: 28-pl

Content: No Waiver. Failure or delay in exercising any right or remedy under this Agreement shall not constitute a waiver of such (or any other) right or remedy.

Source: 4-pl

Content: And the Distributor has violated or caused Google to violate any Anti-Bribery Laws (as defined in Clause 8.5) or that such a violation is reasonably likely to occur,

Source: 4-pl

Content: And the Distributor has violated or caused Google to violate any Anti-Bribery Laws (as defined in Clause 8.5) or that such a violation is reasonably likely to occur,

Source: 28-pl

FINAL ANSWER: This Agreement is governed by English law.

SOURCES: 28-pl

23
A unity agenda for the nation.
We can do this.
My fellow Americans|tonight , we have gathered in a sacred space|the citadel of our democracy.
In this Capitol, generation after generation, Americans have debated great questions amid great strife, and have done
great things.
We have fought for freedom, expanded liberty, defeated totalitarianism and terror.
And built the strongest, freest, and most prosperous nation the world has ever known.
Now is the hour.
Our moment of responsibility.
Our test of resolve and conscience, of history itself.
It is in this moment that our character is formed. Our purpose is found. Our future is forged.
Well I know this nation.

Source: 34-pl

Figure 5: Langchain example prompt for QA with sources.

You are a MyScale expert. Given an input question, first create a syntactically correct MyScale query to run, then look
at the results of the query and return the answer to the input question.
MyScale queries has a vector distance function called DISTANCE(column, array) to compute relevance to the user's question
and sort the feature array column by the relevance.
When the query is asking for {top:k} closest row, you have to use this distance function to calculate distance to entity's
array on vector column and order by the distance to retrieve relevant rows.
*NOTICE*: DISTANCE(column, array) only accept an array column as its first argument and a NeuralArray(entity) as its second
argument. You also need a user defined function called NeuralArray(entity) to retrieve the entity's array.
Unless the user specifies in the question a specific number of examples to obtain, query for at most {top:k} results using
the LIMIT clause as per MyScale. You should only order according to the distance function.
Never query for all columns from a table. You must query only the columns that are needed to answer the question. Wrap
each column name in double quotes (") to denote them as delimited identifiers.
Pay attention to use today() function to get the current date, if the question involves "today". ORDER BY clause should
always be after WHERE clause. DO NOT add semicolon to the end of SQL. Pay attention to the comment in table schema.

Use the following format:

```
[1] [table info] [2]
[3] [table info] [4]
[7] SQLQuery: [input] [8]
```

Here are some examples:

```
[1] CREATE TABLE "ChatPaper" (
[2] abstract String,
[3] id String,
[4] vector Array(Float32),
[5] ) ENGINE = ReplicatedReplacingMergeTree()
[6] ORDER BY id
[7] PRIMARY KEY id
[8] Question: What is Feature Pyramid Network?
[10] NeuralArray(PaperRank contribution)) LIMIT {top:k}
[11] Let's begin:
```

Figure 6: Langchain example prompt for SQL querying using MyScale.
Changes for 2022 that can affect tax refund:

1. **Changes in the number of dependents**: Employment or self-employment income and divorce, among other factors, may affect your tax-filing status and refund. No additional stimulus payments for 2022 so taxpayers should not expect to get an additional payment.
2. **Some tax credits return to 2019 levels**: This means that taxpayers will likely receive a significantly smaller refund compared with the previous tax year. Changes include amounts for the Child Tax Credit (CTC), the Earned Income Tax Credit (EITC) and the Child and Dependent Care Credit will revert to pre-COVID levels.
3. **For 2022, the CTC is worth $2,000 for each qualifying child**: A child must be under age 17 at the end of 2022 to be a qualifying child. For the EITC, eligible taxpayers with no children will get $560 for the 2022 tax year. The Child and Dependent Care Credit returns to a maximum of $2,100 in 2022.
4. **No above-the-line charitable deductions**: During COVID, taxpayers were able to take up to a $600 charitable donation tax deduction on their tax returns. However, for tax year 2022, taxpayers who don’t itemize and who take the standard deduction, won’t be able to deduct their charitable contributions.
5. **More people may be eligible for the Premium Tax Credit**: For tax year 2022, taxpayers may qualify for temporarily expanded eligibility for the premium tax credit.
6. **Eligibility rules changed to claim a tax credit for clean vehicles**: Review the changes under the Inflation Reduction Act of 2022 to qualify for a Clean Vehicle Credit.

Figure 8: LlamaIndex example prompt for IRS chatbot guidelines.
F SOFTWARE DEPENDENCIES

DSPy has the following core software requirements:

- **Python**: Version 3.9 or higher.
- **HuggingFace Text Generation Inference server**: Used for hosting language models from the HuggingFace transformers library.
- **OpenAI connectivity**: Required for accessing GPT models.
- **ColBERTv2 retrieval server**: Serves as the default search index retriever.

For optimal performance:

- Access to GPUs is recommended for hosting the language and retrieval models.
- DSPy experiments specifically relied on **NVIDIA A100-SXM GPUs** with 80 GiBs memory.

For model fine-tuning, the following package versions were used:

- `datasets=2.14.5`
- `transformers=4.32.0`
- `peft=0.5.0`
- `trl=0.7.1`

G MODULES

G.1 Predict

```python
class Predict(dspy.Module):
    def __init__(self, signature, **config):
        self.signature = dspy.Signature(signature)
        self.config = config
        # Module Parameters.
        self.lm = dspy.ParameterLM(None)  # use the default LM
        self.demonstrations = dspy.ParameterDemonstrations([])

def forward(self, **kwargs):
    lm = get_the_right_lm(self.lm, kwargs)
    signature = get_the_right_signature(self.signature, kwargs)
    demonstrations = get_the_right_demonstrations(self.demonstrations, kwargs)

    prompt = signature(demos=self.demos, **kwargs)
    completions = lm.generate(prompt, **self.config)
    prediction = Prediction.from_completions(completions, signature=signature)

    if dsp.settings.compiling is not None:
        trace = dict(predictor=self, inputs=kwargs, outputs=prediction)
        dsp.settings.traces.append(trace)
    return prediction
```

G.2 Chain of Thought

```python
class ChainOfThought(dspy.Module):
    def __init__(self, signature):
        # Modify signature from '*inputs -> *outputs' to '*inputs -> rationale, *outputs'.
        rationale_field = dspy.OutputField(prefix="Reasoning: Let's think step by step.")
        signature = dspy.Signature(signature).prepend_output_field(rationale_field)

        # Declare a sub-module with the modified signature.
        self.predict = dspy.Predict(self.signature)

def forward(self, **kwargs):
    # Just forward the inputs to the sub-module.
    return self.predict(**kwargs)
```
**H TELEPROMPTERS**

**H.1 BootstrapFewShot**

```python
class SimplifiedBootstrapFewShot(Teleprompter):
    def __init__(self, metric=None):
        self.metric = metric

    def compile(self, student, trainset, teacher=None):
        teacher = teacher if teacher is not None else student
        compiled_program = student.deepcopy()

        # Step 1: Prepare mappings between student and teacher Predict modules.
        # Note: other modules will rely on Predict internally.
        assert student_and_teacher_have_compatible_predict_modules(student, teacher)
        name2predictor, predictor2name = map_predictors_recursively(student, teacher)

        # Step 2: Bootstrap traces for each Predict module.
        # We'll loop over the training set. We'll try each example once for simplicity.
        for example in trainset:
            if we_found_enough_bootstrapped_demos(): break

            # turn on compiling mode which will allow us to keep track of the traces
            with dspy.setting.context(compiling=True):
                # run the teacher program on the example, and get its final prediction
                # note that compiling=True may affect the internal behavior here
                prediction = teacher(**example.inputs())

                # get the trace of the all internal Predict calls from teacher program
                predicted_traces = dspy.settings.trace

                # if the prediction is valid, add the example to the traces
                if self.metric(example, prediction, predicted_traces):
                    for predictor, inputs, outputs in predicted_traces:
                        d = dspy.Example(automated=True, **inputs, **outputs)
                        predictor_name = self.predictor2name[id(predictor)]
                        compiled_program[predictor_name].demonstrations.append(d)

        return compiled_program
```

**H.2 BootstrapFewShotWithRandomSearch**

```python
class SimplifiedBootstrapFewShotWithRandomSearch(Teleprompter):
    def __init__(self, metric=None, trials=16):
        self.metric = metric
        self.trials = trials

    def compile(self, student, *, teacher=None, trainset, valset=None):
        valset = trainset if valset is None else valset

        candidates = []
        for seed in range(self.trials):
            # Create a new basic bootstrap few-shot program.
            shuffled_trainset = shuffle(trainset, seed=seed)
            tp = BootstrapFewShot(metric=metric, max_bootstrap_demos=random_size())
            candidate_program = tp.compile(student, shuffled_trainset, teacher)

            # Step 2: Evaluate the generated candidate program.
            score = evaluate_program(candidate_program, self.metric, valset)
            candidates.append((score, candidate_program))

        return max(candidates, key=lambda x: x[0])[1]
```

---

27
I  EXAMPLES OF THE PROMPTS AUTOMATICALLY GENERATED BY DSPY

We include three example prompts bootstrapped by DSPy for the llama2-13b-chat experiments in this paper in Figures 9, 10, and 11. These include the prompt automatically generated by DSPy for GSM8K vanilla program compiled with bootstrap×2, the GSM8K CoT program compiled with bootstrap, and the second-hop generate_query prompt for HotPotQA’s multihop program. All of these, particularly their demonstrations’ labels and their selection, are generated by DSPy automatically using llama2-13b-chat. Consider the first one, i.e. Figures 9. This shows that, even though the program (vanilla) does not induce a chain of thought, the LM effectively leverages the nature of the GSM8K metric (checking for the final number for evaluation) and bootstraps its own reasoning-based demonstration, which makes the program effective at test time. Next, consider the second-hop query prompt bootstrapped by DSPy using llama2-13b-chat, shown in Figure 11. We hypothesize this prompt works effectively because it demonstrates the process of generating search queries for two complex questions by producing two very different types of queries: a simplified question ("When was the first...?") and a keyword query ("Jeremy Paxman birth year"). This might allow the LM to resolve different questions at test-time more effectively.
Given the fields 'question', produce the fields 'answer'.

---

Follow the following format.

Question: $(question)$
Answer: $(answer)$

---

Question: Jimmy and Irene go shopping for clothes on a Tuesday, where senior citizens get a 10% discount on their purchases. Jimmy picks out 3 shorts from the $15 rack. Irene grabs 5 shirts from the $17 rack. How much money do they give to the cashier?
Answer: Jimmy picks out 3 shorts at $15 each = $45. Irene grabs 5 shirts at $17 each = $85. Total cost = $45 + $85 = $130. Since senior citizens get a 10% discount, they will pay 10% of $130 = $13. So, they will give the cashier $130 - $13 = $117.

---

Question:

Figure 9: Copy of the prompt automatically generated by DSPy for GSM8K Llama2-13b-chat vanilla program compiled with bootstrap $\times 2$.

---

Given the fields 'question', produce the fields 'answer'.

---

Follow the following format.

Question: $(question)$
Reasoning: Let's think step by step in order to produce the answer. We ...
Answer: $(answer)$

---

Question: Mark is baking bread. He has to let it rise for 120 minutes twice. He also needs to spend 10 minutes kneading it and 30 minutes baking it. How many minutes does it take Mark to finish making the bread?
Reasoning: Let's think step by step in order to find out how many minutes it takes Mark to finish making the bread. We know that he needs to let it rise for 120 minutes twice, so that's 240 minutes. Then, he needs to spend 10 minutes kneading it and 30 minutes baking it. So, in total, it will take Mark 240 + 10 + 30 = 280 minutes to finish making the bread.
Answer: 280 ---

---

Question: Ben has $2000 for his business operations costs. He orders goods from his supplier and writes them a cheque for $600. His debtor pays him $800 from the purchases they had made on credit. Mr. Ben then decides to do equipment maintenance and spends $1200 on the whole operation. How much money is Mr. Ben remaining with?
Reasoning: Let's think step by step in order to find out how much money Mr. Ben is remaining with. We know that he had $2000 to begin with, and he spent $600 on goods from his supplier, so he has $2000 - $600 = $1400 left. Then, his debtor paid him $800, so he has $1400 + $800 = $2200. Finally, he spent $1200 on equipment maintenance, so he has $2200 - $1200 = $1000 left.
Answer: $1000 ---

---

... several other demonstrations here ...

---

Question:

Figure 10: Shortened copy of the prompt automatically generated by DSPy for GSM8K Llama2-13b-chat CoT program compiled with bootstrap.
Given the fields 'context', 'question', produce the fields 'search_query'.

Follow the following format.

Context: ${context}
Question: ${question}
Reasoning: Let's think step by step in order to ${produce the search_query}. We ...
Search Query: ${search_query}

---

Context:

[1] Twilight (novel series) | Twilight is a series of four vampire-themed fantasy romance novels by American author Stephenie Meyer. ...
[2] Harper Connelly Mysteries | The Harper Connelly Mysteries is a series of fantasy mystery novels written by Charlaine Harris, and first published in 2005. ...
[3] The Dark Heroine | The Dark Heroine is a series of vampire-themed fantasy romance novels written by English author Abigail Gibbs, published by HarperCollins in 2012. ...

Question: In which year was the first of the vampire-themed fantasy romance novels for which The Twilight Saga: The Official Illustrated Guide serves as a spin-off encyclopedic reference book first published?
Reasoning: Let's think step by step in order to determine the year the first of the vampire-themed fantasy romance novels was first published. ...
Search Query: When was the first of the vampire-themed fantasy romance novels published?

---

Context:

[1] The Victorians | The Victorians - Their Story In Pictures is a 2009 British documentary series which focuses on Victorian art and culture. ....
[2] The Caxtons | The Caxtons: A Family Picture is an 1849 Victorian novel by Edward Bulwer-Lytton that was popular in its time.

Question: The Victorians - Their Story In Pictures is a documentary series written by an author born in what year?
Reasoning: Let's think step by step in order to produce the search query. We know that the documentary series is about Victorian art and culture, and it was written and presented by Jeremy Paxman. Therefore, we need to find the year in which Jeremy Paxman was born.
Search Query: Jeremy Paxman birth year

---

Context:
J Future Work Directions

As we discuss in Sec 2, this work is inspired by the role that abstractions and libraries for deep learning played in the development of the field. Our goal for DSPy is to recreate that revolution for what we call foundation model programming. To this end, we have started working on several follow-up directions that expand and improve DSPy.

At a high level, we see highly promising directions in: (1) the implementation of new modules (corresponding to various new prompting techniques), (2) the systematic evaluation of a larger number of advanced programs for open-ended tasks, (3) the addition of various automatic LM-based metrics for optimization, (4) the addition of methods for better debugging and inspection and (5) for more controlled generation (a la LMQL; Beurer-Kellner et al. 2023), and (6) the development a deeper understanding of the tradeoffs associated with compiling with large vs. small LMs.

Beyond these, one of the key strengths of DSPy is the modularity between programs and teleprompters (optimizers). The discussion in Sec 4 lays out several angles for building more sophisticated teleprompters, which can improve the quality of existing as well as new DSPy programs.

**Better candidate generation.** We envision a more systematic treatment of the variables to optimize, e.g. instructions, demonstrations, string prefixes (e.g., “Let’s think step by step” for chain of thought), and numerical hyperparameters (e.g., the number of passages to retrieve or the number of candidate answers to generate). Future work may benefit from the way variables are defined in Optuna (Akiba et al., 2019) and in the conceptual framework of Dohan et al. (2022) to allow expressing these. Future work may also evaluate methods to generate more diverse or more challenging demonstrations, including more sophisticated logic for bootstrapping (e.g., trying multiple variants of the program) or filtering (e.g., only selecting examples that the zero-shot version of the program cannot solve correctly).

**Better parameter optimization.** More sophisticated programs will likely benefit from more advanced optimization strategies. For instance, the Optuna teleprompter in DSPy may allow Bayesian optimizations that exceed the quality achieved with random search. More research is required to understand the tradeoffs of cost and quality in this case. Going a step beyond this, LMs can take a more active role in optimization, beyond bootstrapping demonstrations. For instance, they can revise instructions based on categories of common failures or attempt to simulate a notion of gradients over multi-stage programs.

**Better higher-order program optimizations.** All programs in this paper were directed to a prompt (or ensemble of optimized prompts) for each module. It is possible to treat different inputs for each module more dynamically, directing them to more specialized prompts. While DSPy has an initial teleprompter that treats different inputs more dynamically, this space is very large and warrants much exploration. Beyond that, the ability to backtrack to correct mistakes and address LM feedback accordingly is another promising direction.

In all of these, the references presented in Sec 2 may serve as powerful starting points for each individual component (e.g., utilizing LM feedback), which DSPy can generalize and extend to deal with arbitrary LM pipelines of declarative modules.