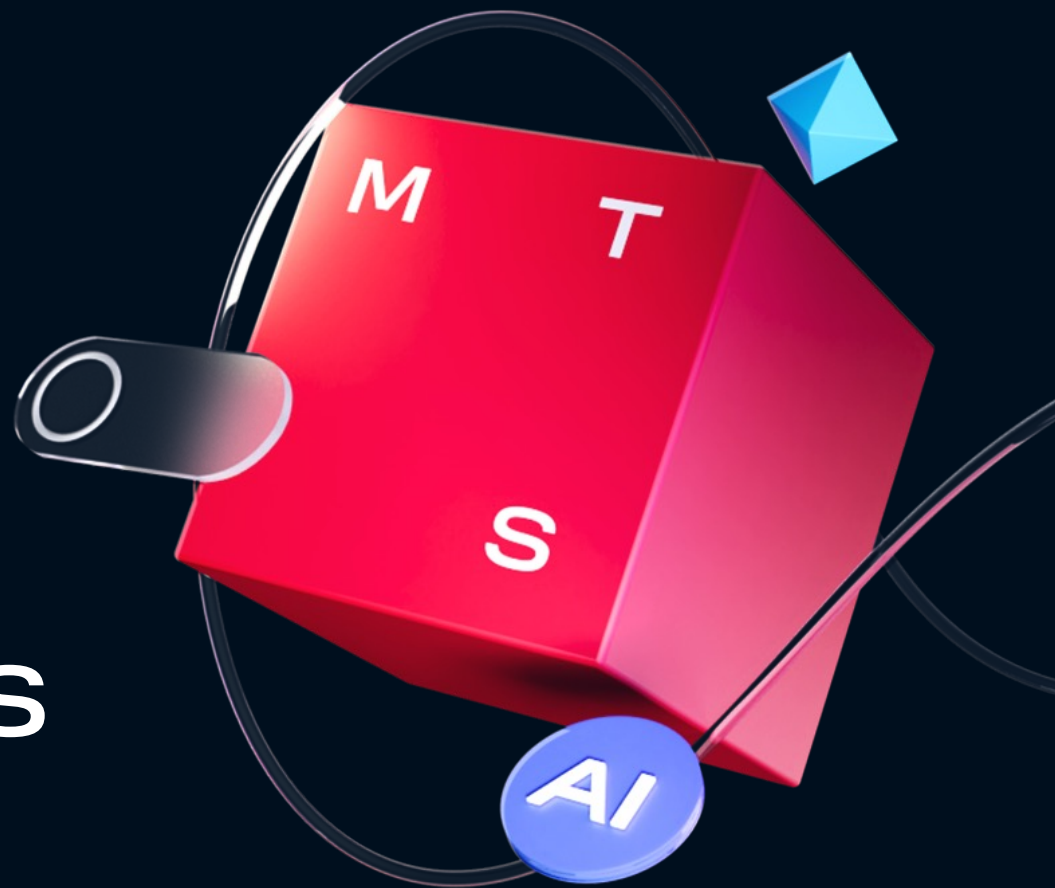




HOW WE BUILD A DEVELOPER'S HELPER

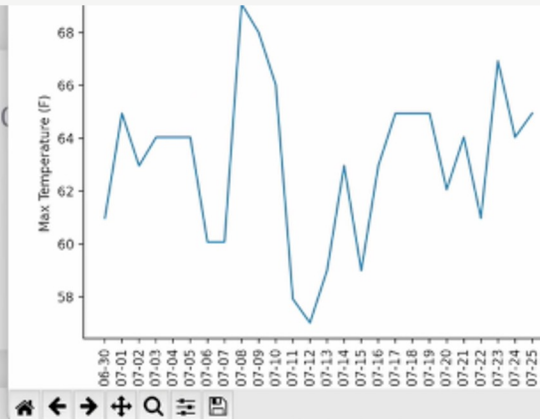
Valentin Malykh

Ноябрь 2023



PLP or NLP for Code

Provide instructions



Now plot the results. Label both axes (y axis is max temperature), rotate the x ticks, and add a title.

Select only the rows with datatype TMAX, and convert from C to F. Also change dates to MM-DD

date	datatype	station
attributes	value	

```
df['date'] =  
pd.to_datetime(df['date'])  
df['date'] =  
df['date'].dt.strftime('%m  
-%d')  
df['value'] = df['value']  
* 1.8 + 32  
  
print(df)
```

```
"""Now plot the results.  
Label both axes (y axis is  
max temperature), rotate  
the x ticks, and add a  
title."""  
import matplotlib.pyplot  
as plt
```

```
plt.plot(df['date'],  
df['value'])  
plt.xlabel('Date')  
plt.ylabel('Max  
Temperature (F)')  
plt.xticks(rotation=90)  
plt.title('Max Temperature  
in San Francisco')  
plt.show()
```

sentiment.ts

write_sql.go

parse_expenses.py

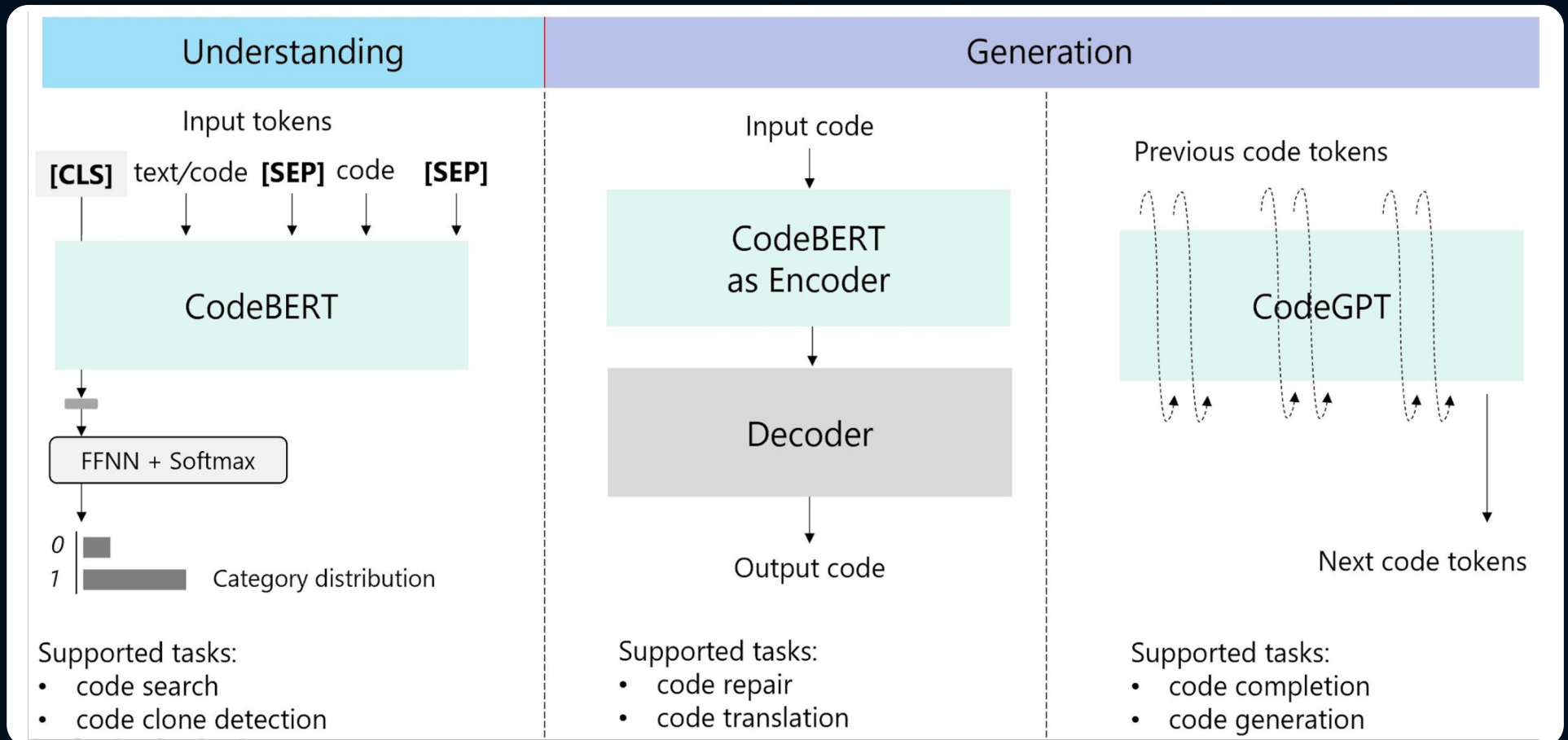
addresses.rb

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

Copilot

Category	Task	Dataset Name	Language	Train/Dev/Test Size	Baselines	Task definition
Code-Code	Clone Detection	BigCloneBench	Java	900K/416K/416K	CodeBERT	Predict semantic equivalence for a pair of codes.
		POJ-104	C/C++	32K/8K/12K		Retrieve semantically similar codes.
	Defect Detection	Devign	C	21k/2.7k/2.7k		Identify whether a function is vulnerable.
	Cloze Test	CT-all	Python, Java, PHP, JavaScript, Ruby, Go	-/-/176k		Tokens to be predicted come from the entire vocab.
		CT-max/min	Python, Java, PHP, JavaScript, Ruby, Go	-/-/2.6k		Tokens to be predicted come from {max, min}.
	Code Completion	PY150	Python	100k/5k/50k	CodeGPT	Predict following tokens given contexts of codes.
		GitHub Java Corpus	Java	13k/7k/8k		
	Code Repair	Bugs2Fix	Java	98K/12K/12K	Encoder-Decoder	Automatically refine codes by fixing bugs.
Code Translation	CodeTrans	Java-C#	10K/0.5K/1K	Translate the codes from one programming language to another programming language.		
Text-Code	NL Code Search	CodeSearchNet, AdvTest	Python	251K/9.6K/19K	CodeBERT	Given a natural language query as input, find semantically similar codes.
		CodeSearchNet, WebQueryTest	Python	251K/9.6K/1k		Given a pair of natural language and code, predict whether they are relevant or not.
	Text-to-Code Generation	CONCODE	Java	100K/2K/2K	CodeGPT	Given a natural language docstring/comment as input, generate a code.
Code-Text	Code Summarization	CodeSearchNet	Python, Java, PHP, JavaScript, Ruby, Go	908K/45K/53K	Encoder-Decoder	Given a code, generate its natural language docstring/comment.
Text-Text	Documentation Translation	Microsoft Docs	English-Latvian/Danish/Norwegian/Chinese	156K/4K/4K		Translate code documentation between human languages (e.g. En-Zh), intended to test low-resource multi-lingual translation.

Task Types



Business Application

The screenshot displays an IDE interface for a Django project. The main editor shows a Python test file with the following code:

```
create_question(question_text="Past question.", days=-30)
response = self.client.get(reverse('polls:index'))
self.assertQuerysetEqual(
    response.context['latest_question_list'],
    ['<Question: Past question.>']
)

def test_index_view_with_a_future_question(self):
    """
    Questions with a pub_date in the future should
    not be displayed on the index page.
    """
    create_question(question_text="Future question.", days=30)
    response = self.client.get(reverse('polls:index'))
    self.assertContains(response, "No polls are available",
                        status_code=200)
    self.assertQuerysetEqual(response.context['latest_question_list'], [])

def test_index_view_with_future_question_and_past_question(self):
    """
    Even if both past and future questions exist,
    only the past question should be displayed.
    """
    create_question(question_text="Past question.", days=-30)
    create_question(question_text="Future question.", days=30)
    response = self.client.get(reverse('polls:index'))
    self.assertQuerysetEqual(
        response.context['latest_question_list'],
        ['<Question: Past question.>']
    )
```

A search overlay is active, showing the following results:

- Search Everywhere: Include non-project items (Double ⇧) ✖
- Search: result
- Classes (⌘O): ResultsView (polls.views)
- Files (⇧O): results.html (djtp_first_steps/polls/templates/polls)
- Symbols (⇧⌘O): result (FileReader (HTML5.js)), result (StdSuites.AppleScript_Suite), result (e (Python 2.7.8 virtualenv at ~/django_p27)), result (event (Python 2.7.8 virtualenv at ~/django_p27)), result (event (Python 2.7.8 virtualenv at ~/django_p27)), ... more
- Actions (⇧⌘A): View Offline Inspection Results... Code
- Import Test Results

The Database tool window on the right shows the following structure:

- Django default
- tables 13
 - auth_group
 - auth_group_permissions
 - auth_permission
 - auth_user
 - auth_user_groups
 - auth_user_user_permissions
 - django_admin_log
 - id INTEGER
 - action_time TEXT
 - object_id TEXT
 - object_repr TEXT
 - action_flag INTEGER
 - change_message TEXT
 - content_type_id INTEGER
 - user_id INTEGER
 - <unnamed> (id)
 - #FAKE_django_admin_log
 - #FAKE_django_admin_log
 - django_admin_log_417f1
 - django_admin_log_e8701
 - django_content_type
 - django_migrations

The Debug console at the bottom shows the following variables:

- longMessage = {bool} False
- maxDiff = {int} 640
- reset_sequences = {bool} False
- serialized_rollback = {bool} False
- startTime = {datetime} 2015-10-09 11:38:35.521452

The status bar at the bottom indicates: Tests Failed: 4 passed, 3 failed (4 minutes ago). The system tray shows the time 34:9, locale Lf, encoding UTF-8, and the current branch Git: master.

Code Completion

```
public class CodeCompletion {  
  
    private void completion() {  
        Square square = new Square();  
        String string = "";  
        List<String> strings;
```

square. |

}	m getId()	String
	m draw()	void
private s	m equals(Object obj)	boolean
	m hashCode()	int
private s	m toString()	String
	m getClass()	Class<? extends Square>
private s	m notify()	void
	m notifyAll()	void
}	m wait()	void
	m wait(long timeoutMillis)	void
	m wait(long timeoutMillis, int nanos)	void

Code Completion Dataset

There is no common benchmark

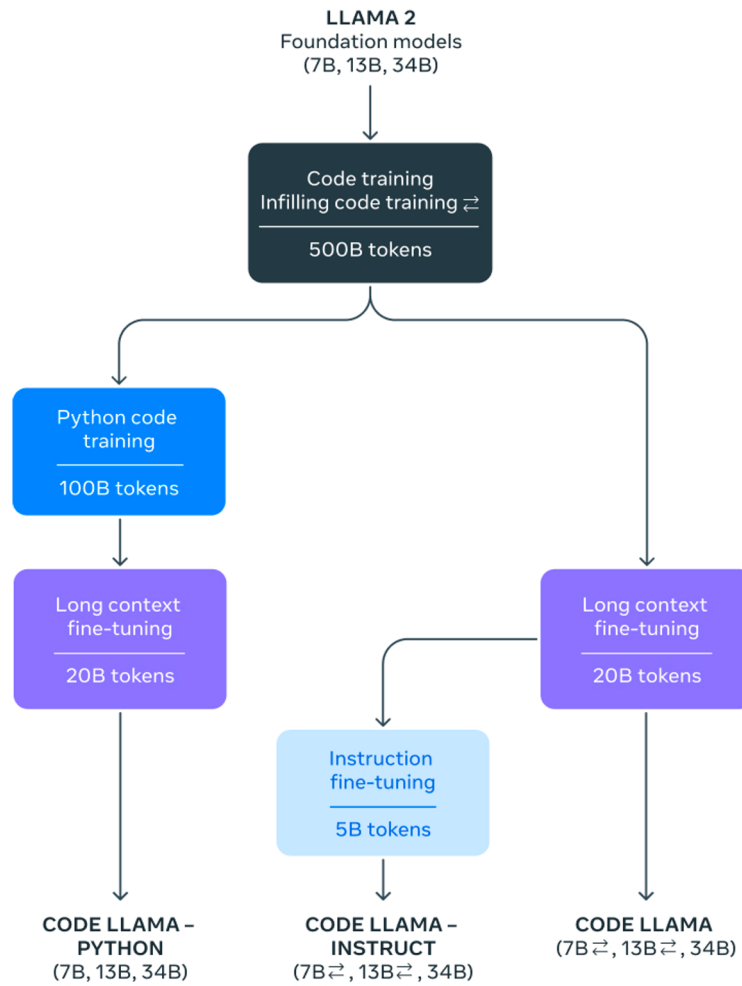
We use our internal code set

- Private
- 1k snippets

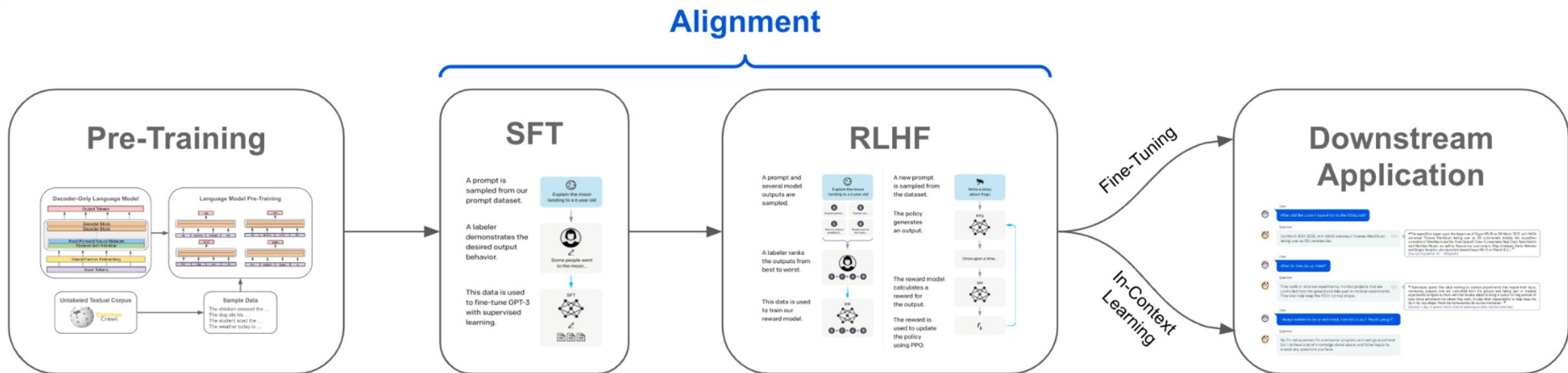
```
public class CodeCompletion {  
  
    private void completion() {  
        Square square = new Square();  
        String string = "";  
        List<String> strings;  
  
        square.|  
    }  
  
    private s  
  
    private s  
  
    private s  
}
```

m	getId()	String
m	draw()	void
m	equals(Object obj)	boolean
m	hashCode()	int
m	toString()	String
m	getClass()	Class<? extends Square>
m	notify()	void
m	notifyAll()	void
m	wait()	void
m	wait(long timeoutMillis)	void
m	wait(long timeoutMillis, int nanos)	void

Model: Code Llama



Model: Training

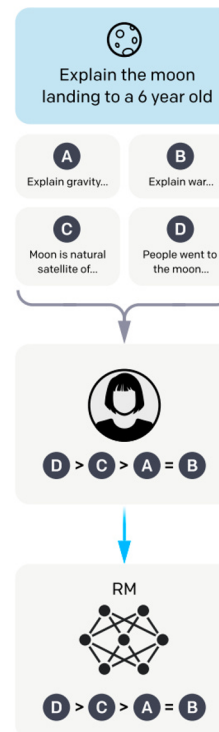


Model: Reinforcement Learning with Human Feedback

Step 2

**Collect comparison data,
and train a reward model.**

A prompt and
several model
outputs are
sampled.



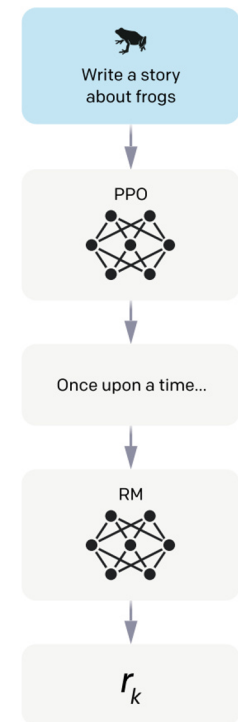
A labeler ranks
the outputs from
best to worst.

This data is used
to train our
reward model.

Step 3

**Optimize a policy against
the reward model using
reinforcement learning.**

A new prompt
is sampled from
the dataset.



The policy
generates an
output.

The reward model
calculates a
reward for
the output.

The reward is
used to update
the policy
using PPO.

Code Completion: Approach

01

(Now)
We work with
Python only

02

We use Jedi

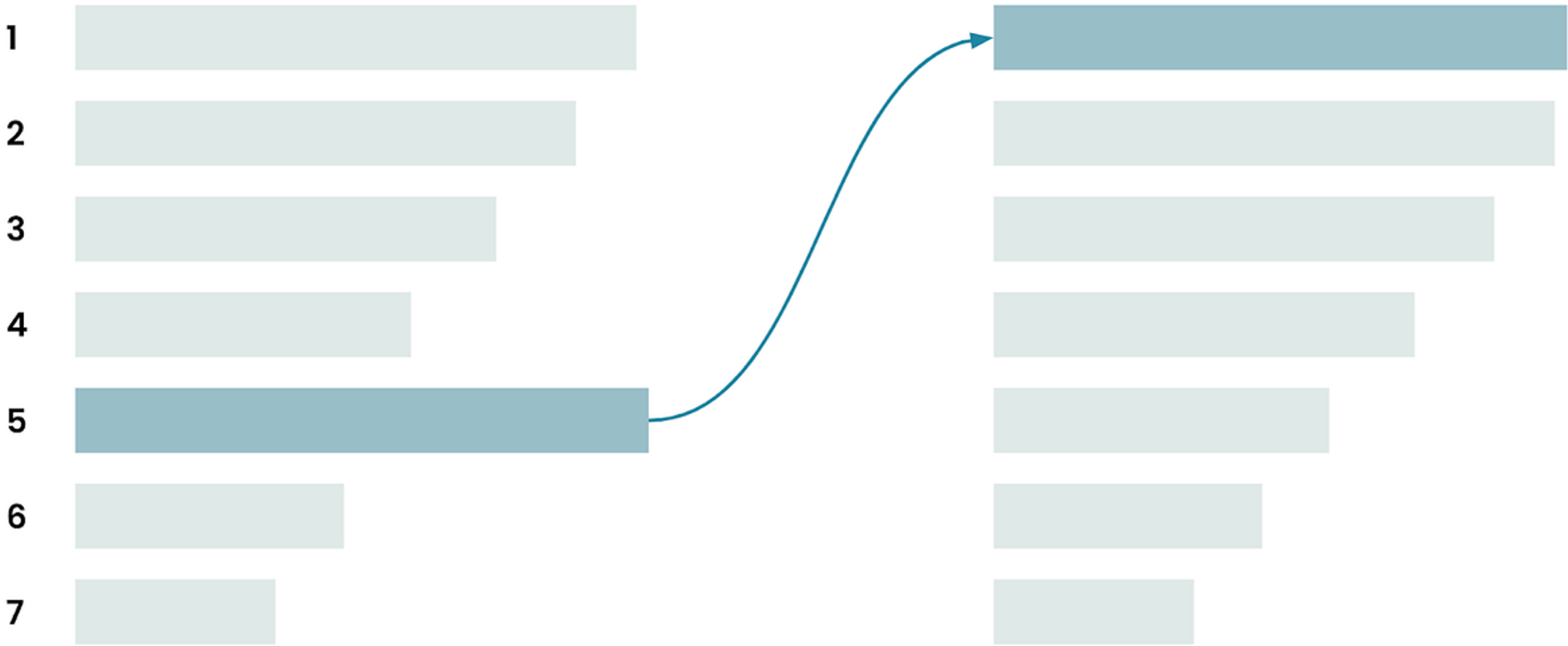


03

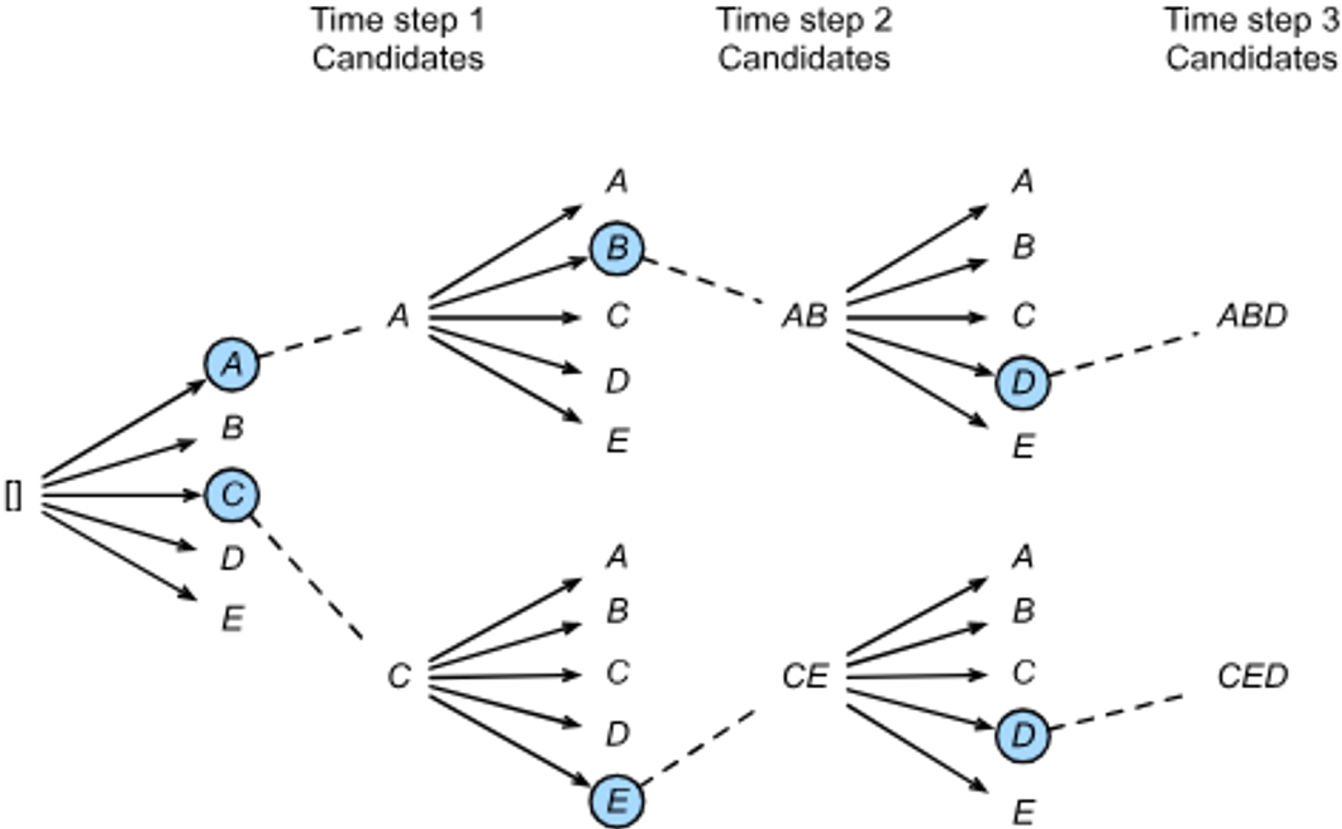
We use Jedi
output to:

- Rerank the candidates
- Constrain beam search

Code Completion: Reranking



Code Completion: Beam Search



Code Completion: Results

	mts-sdk	mts-sdk jedi has the target
codegen25-7b-multi (generation)	0.3970	0.3730
codegen25-7b-multi (jedi ranking)	0.3190	0.3670
codegen25-7b-multi (jedi ranking) alt.	0.3290	0.3795
codellama-7b-base (jedi ranking)	0.3926	0.4518
codellama-7b-base (jedi ranking) alt.	0.3560	0.4096
codellama-7b-base (generation)	0.3926	0.3614
codellama-7b-base (constrained beam search)	0.3979	0.3674

Code Generation

```
parse_expenses.py write_sql.go sentiment.ts addresses.rb  
1 import datetime  
2  
3 def parse_expenses(expenses_string):  
4     """Parse the list of expenses and return the list of triples (date, va  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
```


Code Generation: Dataset

HumanEval-X Benchmark

Stats: 820 handcrafted samples with test cases covering C++, Java, JavaScript, Go and Python.

Metrics: Pass@ k , Pass@ k_{π} (evaluating multilingual functional correctness)

Tasks: Generation: (1) (2) → (3) (Test: (1) (2) (3) (4))

Translation: (1) (3) (1') → (3') (Test: (1') (2') (3') (4'))

<pre>from typing import List def has_close_elements(numbers: List[float], threshold: float) -> bool:</pre>	Declaration (1)	<pre>import java.util.*; import java.lang.*; class Solution {</pre>	(1')
<pre> """ Check if in given list of numbers, are any two numbers closer to each other than given threshold. ... """</pre>	Docstring (2)	<pre> /**Check if in given list of numbers, are any two numbers closer to each other than given threshold. ...*/</pre>	(2')
<pre> for idx, elem in enumerate(numbers): for idx2, elem2 in enumerate(numbers): if idx != idx2: distance = abs(elem - elem2) if distance < threshold: return True return False</pre>	Solution (3)	<pre> List<Double> numbers, double threshold) { for (int i = 0; i < numbers.size(); i++) { for (int j = i + 1; j < numbers.size(); j++) { double distance = Math.abs(numbers.get(i) - numbers.get(j)); if (distance < threshold) return true;}} return false;}}</pre>	(3')
<pre>def check(candidate): assert candidate([1.0, 2.0, 5.9, 4.0, 5.0], 0.95) == True ... check(has_close_elements)</pre>	Test (4)	<pre>public class Main { public static void main(String[] args) { Solution s = new Solution(); List<Boolean> correct = Arrays.asList(s.has_close_elements(Arrays.asList(1.0, 2.0, 5.9, 4.0, 5.0), 0.95), ...}</pre>	(4')

Python (Problem 0)

Java (Problem 0)

Code Generation: Results

Model	Params	Python	C++	Java	JavaScript	Go
CodeGeeX	13B	22.89	17.06	20.04	17.59	14.43
CodeGen	16B	19.22	18.05	14.95	18.40	13.03
Phind/Phind-CodeLlama-34B-v2	34B	72.56	18.90	-	22.5	59.70
CodeLlama-34b-Instruct	34B	40.20	00.60	-	18.29	32.90
codegen25-7b-multi	8B	31.09	00.60	-	00.60	15.80
CodeLlama-7b-Instruct	7B	35.30	00.00	-	23.1	22.56
CodeFuse-13B-SFT	13B	37.10	19.51	26.22	31.71	24.39
CodeLlama-7b-Instruct + code solution filtration	7B	42.50	14.30	21.80	26.50	19.30

Conclusion

01

We can use techniques from NLP in AI4Code

02

We can improve the existing models with a bit of imagination

03

We need your help with that



ВАЛЕНТИН МАЛЫХ

Ведущий разработчик-
исследователь MTS AI

My contacts

My channel

@kvasovamarina



MTS AI — один из лидеров в области искусственного интеллекта в России

>80 миллионов человек

пользуются нашими решениями каждый день

Более
200
экспертов

Реализовано более
20
проектов

Самый мощный в телекоме суперкомпьютер

Участие в конференциях и победы в международных конкурсах по ИИ

Статьи в научных журналах уровня Tier 1

Автоматизировали клиентский сервис МТС с помощью чат-бота



Создали Gen AI помощника для написания кода и чат-бота на основе LLM

Сообщество

30+
партнеров

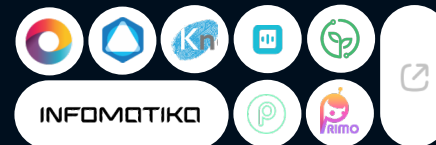
800
стартапов

Помогаем прогнозировать доходность инвестиций в телеком-инфраструктуру с помощью ИИ

Научили KION пропускать титры и определять место для рекламы

8+

Портфельных компаний



Области компетенций MTS AI

CV:

Компьютерное зрение

- Детекция объектов на видео: людей, транспорт, поиск изображений по критериям
- Распознавание лиц
- Контроль пересечения линий и выбранных зон

NLP:

Обработка естественных языков

- Голосовые и текстовые боты для автоматизации общения с клиентами
- Озвучивание текстов синтезированным голосом
- Распознавание речи

Gen AI:

Генеративные сети

- Персонализированные чат-боты
- Генерация текстов, видео и изображений по описанию

RPA:

Роботизация процессов

- Проверка отчетов
- Оформление документов
- Мониторинг почты
- Обработка списаний
- Выставление счетов

OCR:

Распознавание символов

- Распознавание рукописного текста
- Распознавание бумажных документов
- Оцифровка документов

Проекты MTS AI

Смарти

Бот обслуживает
80 млн абонентов MTC
в чатах и на сайте

ИИ-оператор (voice bot)

Голосовой бот, на базе Audiogram,
консультирует абонентов
в контакт-центре MTC

Виртуальный ассистент

Аватар MTC,
объединяющий
сервисы в экосистеме



KION

Автоматическая разметка
контента: поиск места для
вставки рекламы, пропуск
титров, генерация постеров

Цифровой портрет телекома

Имитационная
ML-модель для
прогнозирования
доходности бизнеса

Цифровая маркировка

Генерация
и распознавание
аудиометок для
использования
в маркетинговых
материалах

MTC видеонаблюдение

Сервис облачного видео-
наблюдения и аналитики
для среднего и малого
бизнеса, созданный
на базе TenVision MTS AI



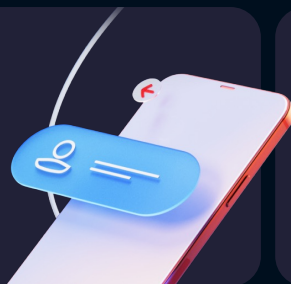
ГОЛЬФСТРИМ

ОХРАННЫЕ СИСТЕМЫ

Интеграция облачного
видеонаблюдения
и аналитики в охранные
сервисы. Разработка
агента для камер

Защитник/ секретарь

Функция расшифровки
спам-звонков на базе
Audiogram. Абоненту
приходит СМС с текстом
звонка и категорий



Строки

Озвучивание книг
синтезированным
голосом с помощью
Audiogram, улучшение
качества синтеза

Генеративные сети

Большая языковая модель
на базе open-source-моделей
и своих разработок. Сервис
генерации и втдополнения
кода на основе Gen AI

Более 100 пилотов

Аналитика коммуникация
для банков, телекома
и e-commerce, видеоаналитика
для розничных магазинов