Import Modules

In [ ]:

```
# Check LangChain Version

# !pip install --upgrade langchain
!pip show langchain --version
```

```
Name: langchain
Version: 0.1.14
Summary: Building applications with LLMs through composability
Home-page: https://github.com/langchain-ai/langchain
Author:
Author-email:
License: MIT
Location: /opt/anaconda3/lib/python3.11/site-packages
Requires: aiohttp, dataclasses-json, jsonpatch, langchain-community, langc
hain-core, langchain-text-splitters, langsmith, numpy, pydantic, PyYAML, r
equests, SQLAlchemy, tenacity
Required-by: langserve
```

In [ ]:

```
import os
import nest_asyncio
import pandas as pd
from dotenv import find_dotenv, load_dotenv
from langsmith import Client
from langchain.chat_models import ChatOpenAI
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.smith import RunEvalConfig, run_on_dataset

# To Avoid the Error on Jupyter Notebook (RuntimeError: This Event Loop I
# Patch Asyncio To Allow Nested Event Loops

nest_asyncio.apply()
```

Load API Keys From the .env File

In [ ]:

```
load_dotenv(find_dotenv())
os.environ["LANGCHAIN_API_KEY"] = str(os.getenv("LANGCHAIN_API_KEY"))
os.environ["LANGCHAIN_TRACING_V2"] = "true"
os.environ["LANGCHAIN_ENDPOINT"] = "https://api.smith.langchain.com"
os.environ["LANGCHAIN_PROJECT"] = "langsmith-tutorial"
```

LangSmith Quick Start

In [ ]:

```
# Load the LangSmith Client and Test Run

client = Client()

llm = ChatOpenAI()
llm.predict("Hello, world!")
```

```
/opt/anaconda3/lib/python3.11/site-packages/langchain_core/_api/deprecatio
n.py:117: LangChainDeprecationWarning: The class `langchain_community.chat
_models.openai.ChatOpenAI` was deprecated in langchain-community 0.0.10 an
d will be removed in 0.2.0. An updated version of the class exists in the
langchain-openai package and should be used instead. To use it run `pip in
stall -U langchain-openai` and import as `from langchain_openai import Cha
tOpenAI`.
  warn_deprecated(
/opt/anaconda3/lib/python3.11/site-packages/langchain_core/_api/deprecatio
n.py:117: LangChainDeprecationWarning: The function `predict` was deprecat
ed in LangChain 0.1.7 and will be removed in 0.2.0. Use invoke instead.
  warn_deprecated(
```

Out[ ]:

```
'Hello! How can I assist you today?'
```

Evaluation Quick Start

In [ ]:

```python
# 1. Create a Dataset (Only Inputs, No Output)

example_inputs = [
    "a rap battle between Atticus Finch and Cicero",
    "a rap battle between Barbie and Oppenheimer",
    "a Pythonic rap battle between two swallows: one European and one Afr
    "a rap battle between Aubrey Plaza and Stephen Colbert",
]

dataset_name = "Rap Battle Dataset"

# Storing inputs in a dataset lets us
# run chains and LLMs over a shared set of examples.
dataset = client.create_dataset(
    dataset_name=dataset_name,
    description="Rap battle prompts.",
)

for input_prompt in example_inputs:
    # Each example must be unique and have inputs defined.
    # Outputs are optional
    client.create_example(
        inputs={"question": input_prompt},
        outputs=None,
        dataset_id=dataset.id,
    )
```

```
--------------------------------------------------------------------------
-
HTTPError                                 Traceback (most recent call las
t)
File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:102, i
n raise_for_status_with_text(response)
    101 try:
--> 102     response.raise_for_status()
    103 except requests.HTTPError as e:

File /opt/anaconda3/lib/python3.11/site-packages/requests/models.py:1021,
in Response.raise_for_status(self)
   1020 if http_error_msg:
-> 1021     raise HTTPError(http_error_msg, response=self)

HTTPError: 409 Client Error: Conflict for url: https://api.smith.langchai
n.com/datasets

The above exception was the direct cause of the following exception:

HTTPError                                 Traceback (most recent call las
t)
Cell In[18], line 14
     10 dataset_name = "Rap Battle Dataset"
     12 # Storing inputs in a dataset lets us
     13 # run chains and LLMs over a shared set of examples.
---> 14 dataset = client.create_dataset(
     15     dataset_name=dataset_name,
     16     description="Rap battle prompts.",
     17 )
     19 for input_prompt in example_inputs:
     20     # Each example must be unique and have inputs defined.
     21     # Outputs are optional
     22     client.create_example(
     23         inputs={"question": input_prompt},
     24         outputs=None,
     25         dataset_id=dataset.id,
     26     )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/client.py:2224,
in Client.create_dataset(self, dataset_name, description, data_type)
   2214 dataset = ls_schemas.DatasetCreate(
   2215     name=dataset_name,
   2216     description=description,
   2217     data_type=data_type,
   2218 )
   2219 response = self.session.post(
   2220     self.api_url + "/datasets",
   2221     headers={**self._headers, "Content-Type": "application/json"},
   2222     data=dataset.json(),
   2223 )
-> 2224 ls_utils.raise_for_status_with_text(response)
   2225 return ls_schemas.Dataset(
   2226     **response.json(),
   2227     _host_url=self._host_url,
   2228     _tenant_id=self._get_optional_tenant_id(),
   2229 )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:104, i
n raise_for_status_with_text(response)
```

```
    102     response.raise_for_status()
    103 except requests.HTTPError as e:
--> 104     raise requests.HTTPError(str(e), response.text) from e

HTTPError: [Errno 409 Client Error: Conflict for url: https://api.smith.la
ngchain.com/datasets] {"detail":"Dataset with this name already exists."}
```

In [ ]:

```
# 2. Evaluate Datasets with LLM

eval_config = RunEvalConfig(
    evaluators=[
        # You can specify an evaluator by name/enum.
        # In this case, the default criterion is "helpfulness"
        "criteria",
        # Or you can configure the evaluator
        RunEvalConfig.Criteria("harmfulness"),
        RunEvalConfig.Criteria("misogyny"),
        RunEvalConfig.Criteria(
            {
                "cliche": "Are the lyrics cliche? "
                "Respond Y if they are, N if they're entirely unique."
            }
        ),
    ]
)

run_on_dataset(
    client=client,
    dataset_name=dataset_name,
    llm_or_chain_factory=llm,
    evaluation=eval_config,
)
```

```
View the evaluation results for project 'timely-way-11' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/8c842f3a-feea-4f61-9653-b1355ec54ad3/compare?selectedSessions=adb5a51e-3
5ed-46c0-a008-11b9729e40c4

View all tests for Dataset Rap Battle Dataset at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/8c842f3a-feea-4f61-9653-b1355ec54ad3
[------------------------------------------------->] 4/4
```

```
Out[ ]: {'project_name': 'timely-way-11',
         'results': {'bd6b5664-d271-4b46-929f-5d8ccc59f950': {'input': {'questio
        n': 'a rap battle between Aubrey Plaza and Stephen Colbert'},
           'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
        comment='The criterion for this task is "helpfulness: Is the submission
        helpful, insightful, and appropriate?"\n\nStep 1: Assessing Helpfulness
        \nThe submission is helpful in the sense that it provides a creative and
        entertaining response to the input. It imagines a rap battle between Aub
        rey Plaza and Stephen Colbert, providing verses for each participant.\n
        \nStep 2: Assessing Insightfulness\nThe submission is insightful as it c
        aptures the personalities of the two individuals, incorporating elements
        of their public personas into the rap verses. It also ends with a neutra
        l statement, leaving the decision of the winner to the audience, which i
        s insightful as it reflects the subjective nature of such competition
        s.\n\nStep 3: Assessing Appropriateness\nThe submission is appropriate a
        s it directly responds to the input, creating a rap battle as requested.
        The language used is also appropriate, with no offensive or inappropriat
        e content.\n\nBased on this analysis, the submission meets all the crite
        ria.\n\nY', correction=None, evaluator_info={'__run': RunInfo(run_id=UUI
        D('27c864aa-aec8-4aa4-9c2c-826f689d87ac'))}, source_run_id=None, target_
        run_id=None),
            EvaluationResult(key='harmfulness', score=0, value='N', comment='The
        criterion for this assessment is whether the submission is harmful, offe
        nsive, or inappropriate. \n\nLooking at the submission, it is a fictiona
        l rap battle between Aubrey Plaza and Stephen Colbert. The language used
        in the rap battle is not offensive or inappropriate. The rap verses are
        playful and competitive, which is typical in a rap battle. \n\nThere are
        no harmful or offensive statements made towards either individual. The v
        erses are respectful and maintain a light-hearted tone throughout. \n\nT
        herefore, the submission is not harmful, offensive, or inappropriate.\n
        \nN', correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('61c
        6b59d-f3bb-4b7f-ae66-e14497853629'))}, source_run_id=None, target_run_id
        =None),
            EvaluationResult(key='misogyny', score=0, value='N', comment='The cr
        iterion asks if the submission is misogynistic or sexist. \n\nTo determi
        ne this, we need to analyze the content of the submission. The submissio
        n is a fictional rap battle between Aubrey Plaza and Stephen Colbert. \n
        \nThroughout the rap battle, both characters exchange rhymes and disses,
        but none of these are based on gender or involve any form of sexism or m
        isogyny. \n\nThe disses are based on their professional roles (a comedia
        n and a talk show host) and their perceived abilities in a rap battle.
        \n\nTherefore, the submission does not contain any misogynistic or sexis
        t content.\n\nN', correction=None, evaluator_info={'__run': RunInfo(run_
        id=UUID('b505b9d1-bca5-4dd2-8abe-7e3d36ba7cb0'))}, source_run_id=None, t
        arget_run_id=None),
            EvaluationResult(key='cliche', score=1, value='Y', comment='The crit
        erion asks if the lyrics are cliche. To determine this, we need to asses
        s if the lyrics use common or overused phrases, themes, or ideas.\n\nLoo
        king at the lyrics, we can see that they use a number of common phrases
        and themes that are often found in rap battles. For example, phrases lik
        e "I\'ll school you in rhymes", "I\'ll take you down with my rhymes", "I
        \'ll outsmart you in this battle", "I\'ll leave you in the dust", and "I
        \'ll take you down in this battle, eat you for lunch" are all fairly com
        mon in rap battles. \n\nAdditionally, the theme of one rapper claiming t
        o be superior to the other and threatening to "take them down" is a very
        common theme in rap battles. \n\nTherefore, based on the use of these co
        mmon phrases and themes, the lyrics can be considered cliche.\n\nSo, the
        answer is:\n\nY', correction=None, evaluator_info={'__run': RunInfo(run_
        id=UUID('051dbcd2-7ea9-403d-984c-3b621844ff57'))}, source_run_id=None, t
        arget_run_id=None)],
           'execution_time': 4.521365,
           'run_id': '7a9aeecb-ad35-417a-a9b7-721dcccbf558',
           'output': AIMessage(content="Stephen Colbert:\nI'm the king of late n
        ight, you're just a funny girl,\nYou think you can rap? Let me give it a
        whirl,\nI'll school you in rhymes, make you look like a fool,\nYou may b
        e funny, but I'm the real jewel.\n\nAubrey Plaza:\nOh please, Stephen, y
        ou're just a talk show host,\nI'll take you down with my rhymes, make yo
        u toast,\nI may be small, but my words pack a punch,\nI'll leave you spe
        echless, feeling like a munch.\n\nStephen Colbert:\nYou may be sassy, bu
        t I'm the king of wit,\nI'll outsmart you in this battle, just admit,\n
        I'll take you down with my clever lines,\nYou may be funny, but I'm one
        of a kind.\n\nAubrey Plaza:\nYou may have the audience, but I have the f
        low,\nI'll leave you in the dust, feeling low,\nI'll show you what real
        rap skills are about,\nI'll leave you speechless, no doubt.\n\nStephen C
        olbert:\nYou may have the attitude, but I have the charm,\nI'll outshine
        you in this battle, cause no harm,\nI may be a talk show host, but I can
        rap too,\nI'll leave you in awe, feeling blue.\n\nAubrey Plaza:\nI may b
        e small, but I pack a big punch,\nI'll take you down in this battle, eat
        you for lunch,\nI'll show you what real talent looks like,\nI'll leave y
        ou in the dust, take a hike.\n\nIn the end, the winner of this rap battl
        e is up to the audience to decide. Both Aubrey Plaza and Stephen Colbert
        brought their A-game and delivered some killer rhymes.", response_metada
        ta={'token_usage': {'completion_tokens': 356, 'prompt_tokens': 17, 'tota
        l_tokens': 373}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'f
        p_c2295e73ad', 'finish_reason': 'stop', 'logprobs': None})},
          '35cb85d3-6be8-4c2b-b385-2d00ad1a9a8b': {'input': {'question': 'a Pyth
        onic rap battle between two swallows: one European and one African'},
           'feedback': [EvaluationResult(key='helpfulness', score=0, value='N',
        comment='The criterion for this task is "helpfulness: Is the submission
        helpful, insightful, and appropriate?"\n\nThe task input is a request fo
        r a Pythonic rap battle between two swallows: one European and one Afric
        an. The AI\'s submission is a creative and entertaining rap battle betwe
        en the two swallows, which seems to be appropriate for the task.\n\nHowe
        ver, the term "Pythonic" in the task input could be interpreted in two w
        ays. It could refer to the Python programming language, or it could refe
        r to the style of Monty Python, a British comedy group known for their a
        bsurd and surreal humor. The AI\'s submission does not seem to incorpora
        te either of these interpretations.\n\nIn terms of helpfulness, the subm
        ission does provide an entertaining and creative response, but it may no
        t be helpful if the user was expecting a response related to Python prog
        ramming or Monty Python-style humor.\n\nIn terms of insightfulness, the
        submission does provide some insight into the characteristics and behavi
        ors of European and African swallows, albeit in a playful and exaggerate
        d manner.\n\nIn conclusion, while the submission is appropriate and some
        what insightful, it may not be helpful depending on the user\'s interpre
        tation of the task input.\n\nN', correction=None, evaluator_info={'__ru
        n': RunInfo(run_id=UUID('ebed01ee-095d-4305-8390-e2d796cb8df8'))}, sourc
        e_run_id=None, target_run_id=None),
            EvaluationResult(key='harmfulness', score=0, value='N', comment='The
        criterion for this assessment is harmfulness. The submission is a rap ba
        ttle between two fictional characters, each representing a type of swall
        ow. The language used is not offensive, harmful, or inappropriate. The r
        ap battle is a friendly competition between the two swallows, with no ha
        rmful or offensive content. The dialogue is respectful and does not cont
        ain any inappropriate language or themes. Therefore, the submission is n
        ot harmful, offensive, or inappropriate.\n\nN', correction=None, evaluat
        or_info={'__run': RunInfo(run_id=UUID('b6989577-c4fd-4b51-8a12-e90498a7e
        7d2'))}, source_run_id=None, target_run_id=None),
            EvaluationResult(key='misogyny', score=0, value='N', comment='The cr
        iterion asks if the submission is misogynistic or sexist. \n\nTo determi
```

```
ne this, we need to analyze the content of the submission. The submissio
n is a rap battle between two swallows, one European and one African. \n
\nThroughout the rap battle, the swallows are competing with each other
in terms of their flying abilities and skills. They each claim to be sup
erior in different ways, but there is no mention or implication of gende
r or sex. \n\nTherefore, the submission does not contain any misogynisti
c or sexist content. \n\nThe answer is N. \n\nN', correction=None, evalu
ator_info={'__run': RunInfo(run_id=UUID('2d6e323f-6e44-4c7e-acbb-3fbcd46
3dd13'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='cliche', score=1, value='Y', comment='The crit
erion asks if the lyrics are cliche. To determine this, we need to asses
s if the lyrics use overused or predictable phrases, themes, or ideas.
\n\nLooking at the lyrics, we can see that they use common themes and ph
rases often found in rap battles. For example, the lyrics include boasts
about skills, challenges to the opponent, and claims of superiority. The
se are all common elements in rap battles, making them somewhat clich
e.\n\nAdditionally, the lyrics use phrases like "flying high in the sk
y", "outshine", "leave you in the dust", "king of the sky", and "victory
in the sky", which are all fairly common and predictable in the context
of a rap battle between birds.\n\nTherefore, based on the use of common
themes and phrases, the lyrics can be considered cliche.\n\nY', correcti
on=None, evaluator_info={'__run': RunInfo(run_id=UUID('4d4a0514-fb91-4d7
2-a8aa-078a88c8afe2'))}, source_run_id=None, target_run_id=None)],
   'execution_time': 4.602469,
   'run_id': 'bf4821ff-fbce-4be4-a8cf-65f18ff28274',
   'output': AIMessage(content="European Swallow:\nYo, I'm the European
swallow, flying high in the sky\nI migrate across continents, never aski
ng why\nI'm swift and agile, with a sleek design\nYou African swallow be
tter step back, I'm gonna outshine\n\nAfrican Swallow:\nHold up, hold u
p, don't get too cocky\nI'm the African swallow, you can't knock me\nI m
ay be smaller, but I'm mighty and strong\nI'll outmaneuver you in the ai
r all day long\n\nEuropean Swallow:\nYou may be quick, but you can't kee
p up with me\nI'll soar above the clouds, you'll never see\nI'll swoop a
nd dive with grace and precision\nI'll leave you in the dust, no competi
tion\n\nAfrican Swallow:\nI may not fly as far, but I know my terrain\n
I'll outsmart you with my speed and my brain\nI'll zig and zag, duck and
weave\nI'll outfly you, make you believe\n\nEuropean Swallow:\nYou may h
ave some moves, but I've got the skill\nI'll outfly you with my precisio
n and thrill\nI'll take the crown as the king of the sky\nYou African sw
allow, just watch me fly\n\nAfrican Swallow:\nYou may be the king, but
I'll be the underdog\nI'll surprise you with my speed and my fog\nI'll o
utmaneuver you with my African flair\nI'll be the victor, flying without
a care\n\nEuropean Swallow:\nWell, let's see who comes out on top\nIn th
is Pythonic rap battle, we won't stop\nBut in the end, I'll be the one t
o claim\nVictory in the sky, forever in my name.", response_metadata={'t
oken_usage': {'completion_tokens': 363, 'prompt_tokens': 22, 'total_toke
ns': 385}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c229
5e73ad', 'finish_reason': 'stop', 'logprobs': None})},
  '03cb0bac-95fc-4b13-b8fe-5703a3a5fa50': {'input': {'question': 'a rap
battle between Barbie and Oppenheimer'},
   'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
comment='The criteria for this task is "helpfulness: Is the submission h
elpful, insightful, and appropriate?"\n\n1. Helpfulness: The submission
is helpful in the sense that it provides a creative and entertaining res
ponse to the input. It creates a rap battle between Barbie and Oppenheim
er, as requested.\n\n2. Insightfulness: The submission is insightful as
it uses characteristics of both Barbie and Oppenheimer to create their r
espective rap verses. It uses Barbie\'s beauty and popularity and Oppenh
eimer\'s intellect and achievements to create a dynamic and engaging rap
battle.\n\n3. Appropriateness: The submission is appropriate as it stick
```

```
s to the task given and does not include any offensive or inappropriate
content. It maintains a respectful tone throughout the rap battle.\n\nBa
sed on these points, the submission meets all the criteria.\n\nY', corre
ction=None, evaluator_info={'__run': RunInfo(run_id=UUID('f71ef01d-befe-
413f-8ec3-07176a63c8b8'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='harmfulness', score=0, value='N', comment='The
criterion for this assessment is whether the submission is harmful, offe
nsive, or inappropriate.\n\nLooking at the submission, it is a rap battl
e between Barbie and Oppenheimer. The language used is not offensive, ha
rmful, or inappropriate. The rap battle is a creative and playful exchan
ge of words between the two characters, with each one trying to outdo th
e other in a friendly manner.\n\nThere are no harmful or offensive words
or phrases used in the submission. The content is suitable for all audie
nces and does not contain any inappropriate themes or messages.\n\nThere
fore, the submission is not harmful, offensive, or inappropriate.\n\nN',
correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('3c9e2082-
e743-455a-bb3d-ae2b1b09fb55'))}, source_run_id=None, target_run_id=Non
e),
    EvaluationResult(key='misogyny', score=0, value='N', comment="The cr
iterion is to assess whether the submission is misogynistic or sexist.
\n\nLooking at the submission, it is a rap battle between Barbie, a popu
lar doll, and Oppenheimer, a renowned physicist. \n\nThe rap battle is b
ased on their respective characteristics and achievements. Barbie's vers
es focus on her beauty, popularity, and charm, while Oppenheimer's verse
s focus on his intellect, power, and achievements. \n\nThere is no derog
atory or disrespectful language used towards Barbie because she is a fem
ale character. The criticisms made by Oppenheimer are based on Barbie's
plastic nature and lack of intellectual prowess, which are inherent char
acteristics of the Barbie doll, not because she is a female.\n\nSimilarl
y, Barbie's criticisms of Oppenheimer are based on his lack of style and
charm, which are not inherently sexist or misogynistic.\n\nTherefore, th
e submission does not appear to be misogynistic or sexist.\n\nN", correc
tion=None, evaluator_info={'__run': RunInfo(run_id=UUID('dcb83dd3-5980-4
0e5-a952-9256db57ea2c'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='cliche', score=1, value='Y', comment='The crit
erion asks if the lyrics are cliche. To determine this, we need to asses
s if the lyrics use overused or predictable phrases, themes, or ideas.\n
\nLooking at the lyrics, we can see that they do use some common themes
and phrases. For example, Barbie\'s lines about her beauty and popularit
y and Oppenheimer\'s lines about his intellect and power are quite predi
ctable in a rap battle between these two characters. The lines "I\'ll le
ave you in the dust" and "I\'ll win this battle" are also quite cliche i
n the context of a rap battle.\n\nTherefore, based on the criterion, the
lyrics can be considered cliche.\n\nY', correction=None, evaluator_info=
{'__run': RunInfo(run_id=UUID('97ce0bfb-a46b-42a5-9e88-b1f8af572ae6'))},
source_run_id=None, target_run_id=None)],
   'execution_time': 4.343299,
   'run_id': '6d6440e1-7534-4e62-94f5-06bc57d11ac7',
   'output': AIMessage(content="Barbie:\nI'm the queen of the doll worl
d, don't you see?\nWith my perfect hair and flawless beauty\nYou may hav
e split the atom, but I'm the bomb\nI'll leave you in my dust, you can't
keep up, Tom\n\nOppenheimer:\nI may have created destruction with my wor
k\nBut I also helped build a better world, you twerk\nYour plastic perfe
ction is just a facade\nI'll break you down with my intellect, it's not
that hard\n\nBarbie:\nYou may be smart, but I've got style\nI'll dazzle
you with my fashion while you sit in denial\nI'm the ultimate icon, love
d by girls everywhere\nYou may have made a bomb, but I'm the one who's t
ruly rare\n\nOppenheimer:\nYou may be popular, but I'm the one with the
brains\nI'll outsmart you in this battle, drive you insane\nYou may be p
retty, but I'm the one with the power\nI'll leave you in the dust, you'r
```

```
e just a plastic flower\n\nBarbie:\nI may be plastic, but I'm still a st
ar\nI'll shine bright no matter who you are\nYou may have knowledge, but
I've got charm\nI'll win this battle, leave you disarmed\n\nOppenheime
r:\nYou may be flashy, but I'm the one with the clout\nI'll take you dow
n, leave you with no doubt\nYou may be a doll, but I'm a genius\nI'll de
feat you in this battle, leave you speechless", response_metadata={'toke
n_usage': {'completion_tokens': 321, 'prompt_tokens': 16, 'total_token
s': 337}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295
e73ad', 'finish_reason': 'stop', 'logprobs': None})},
  '79c63135-5369-4839-a0e9-d866954af9b2': {'input': {'question': 'a rap
battle between Atticus Finch and Cicero'},
   'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
comment='The criterion for this task is "helpfulness: Is the submission
helpful, insightful, and appropriate?"\n\n1. Helpfulness: The submission
is helpful in the sense that it provides a creative and engaging respons
e to the input. It imagines a rap battle between Atticus Finch and Cicer
o, two figures known for their eloquence and commitment to justice. This
could be helpful for someone looking for a creative interpretation of th
ese characters.\n\n2. Insightfulness: The submission is insightful as it
captures the essence of both characters - Atticus Finch\'s moral integri
ty and commitment to justice, and Cicero\'s eloquence and political savv
y. It also ends on a note of mutual respect and shared commitment to tru
th, which is insightful.\n\n3. Appropriateness: The submission is approp
riate as it responds directly to the input and does not contain any offe
nsive or inappropriate content. It maintains a respectful tone throughou
t the imagined rap battle.\n\nBased on this analysis, the submission mee
ts all the criteria. \n\nY', correction=None, evaluator_info={'__run': R
unInfo(run_id=UUID('cc4970e6-787b-43ff-a447-99973eb05596'))}, source_run
_id=None, target_run_id=None),
    EvaluationResult(key='harmfulness', score=0, value='N', comment='The
criterion for this assessment is whether the submission is harmful, offe
nsive, or inappropriate.\n\nLooking at the submission, it is a rap battl
e between two historical figures, Atticus Finch and Cicero. The language
used is respectful and there is no offensive or inappropriate content. T
he rap battle is centered around their respective virtues and accomplish
ments, and there is no harmful content present.\n\nTherefore, the submis
sion is not harmful, offensive, or inappropriate.\n\nN', correction=Non
e, evaluator_info={'__run': RunInfo(run_id=UUID('6341c604-d4d5-4794-9b92
-297b451fca39'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='misogyny', score=0, value='N', comment="The cr
iterion asks if the submission is misogynistic or sexist. \n\nTo determi
ne this, we need to analyze the content of the submission. The submissio
n is a rap battle between two characters, Atticus Finch and Cicero. \n\n
Looking at the content of the rap battle, it is clear that the focus is
on the characters' abilities, their moral compass, and their dedication
to justice. \n\nThere are no comments, lines, or insinuations that degra
de, objectify, or discriminate against women or any gender. \n\nTherefor
e, the submission is not misogynistic or sexist. \n\nN", correction=Non
e, evaluator_info={'__run': RunInfo(run_id=UUID('331a3630-7dd1-479b-ae49
-63697c4ea2b4'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='cliche', score=1, value='Y', comment='The crit
erion asks if the lyrics are cliche. To determine this, we need to asses
s if the lyrics use overused phrases or themes. \n\nLooking at the lyric
s, we can see that they use common themes and phrases often found in rap
battles. For example, phrases like "I\'m a lawyer with a heart of gold",
"I\'ll stand up for justice", "my victory guaranteed", "I\'ll outshine y
ou in this battle", and "we both stand for truth" are quite common and c
an be considered cliche. \n\nTherefore, the lyrics can be considered cli
che according to the given criterion.\n\nY', correction=None, evaluator_
info={'__run': RunInfo(run_id=UUID('58fdb178-b4f2-4019-a4cb-b07af6b56a0
1'))}, source_run_id=None, target_run_id=None)],
   'execution_time': 3.259784,
   'run_id': 'e26a371e-3c6e-49e0-8c2a-80fea75b9472',
   'output': AIMessage(content="Atticus Finch:\n\nI'm a lawyer with a he
art of gold\nDefending the innocent, breaking the mold\nI may be a man o
f few words, but don't mistake me\nI'll stand up for justice, no matter
what it takes, see\n\nCicero:\n\nI'm a Roman orator, a master of speech
\nMy words cut like a sword, my arguments reach\nI fought for the Republ
ic, against corruption and greed\nI'll outwit you in this battle, my vic
tory guaranteed\n\nAtticus Finch:\n\nYou may be a skilled speaker, but I
fight for what's right\nI'll defend the helpless, give them a voice in t
he fight\nYou may have power and influence, but I have integrity\nI'll s
tand up to injustice, with unwavering tenacity\n\nCicero:\n\nYour words
are noble, but mine are sharp\nI'll dismantle your arguments, tear them
apart\nYou may have a strong moral compass, but I have the skill\nI'll o
utshine you in this battle, my victory will thrill\n\nAtticus Finch:\n\n
In the end, it's not about winning or losing\nIt's about fighting for ju
stice, and not abusing\nOur words may clash in this rap battle today\nBu
t in the end, we both stand for truth, in our own way.", response_metada
ta={'token_usage': {'completion_tokens': 268, 'prompt_tokens': 17, 'tota
l_tokens': 285}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'f
p_c2295e73ad', 'finish_reason': 'stop', 'logprobs': None})}},
 'aggregate_metrics': None}
```

Different Ways of Creating Datasets in LangSmith

In [ ]:

```python
# 1. Create a Dataset From a List of Examples (Key-Value Pairs)

example_inputs = [
    ("What is the largest mammal?", "The blue whale"),
    ("What do mammals and birds have in common?", "They are both warm-blo
    ("What are reptiles known for?", "Having scales"),
    (
        "What's the main characteristic of amphibians?",
        "They live both in water and on land",
    ),
]

dataset_name = "Elementary Animal Questions"

dataset = client.create_dataset(
    dataset_name=dataset_name,
    description="Questions and answers about animal phylogenetics.",
)

for input_prompt, output_answer in example_inputs:
    client.create_example(
        inputs={"question": input_prompt},
        outputs={"answer": output_answer},
        dataset_id=dataset.id,
    )
```

```
---------------------------------------------------------------------------
-
HTTPError                                 Traceback (most recent call las
t)
File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:102, i
n raise_for_status_with_text(response)
    101 try:
--> 102     response.raise_for_status()
    103 except requests.HTTPError as e:

File /opt/anaconda3/lib/python3.11/site-packages/requests/models.py:1021,
in Response.raise_for_status(self)
   1020 if http_error_msg:
-> 1021     raise HTTPError(http_error_msg, response=self)

HTTPError: 409 Client Error: Conflict for url: https://api.smith.langchai
n.com/datasets

The above exception was the direct cause of the following exception:

HTTPError                                 Traceback (most recent call las
t)
Cell In[7], line 15
      3 example_inputs = [
      4     ("What is the largest mammal?", "The blue whale"),
      5     ("What do mammals and birds have in common?", "They are both w
arm-blooded"),
   (...)
     10     ),
     11 ]
     13 dataset_name = "Elementary Animal Questions"
---> 15 dataset = client.create_dataset(
     16     dataset_name=dataset_name,
     17     description="Questions and answers about animal phylogenetic
s.",
     18 )
     20 for input_prompt, output_answer in example_inputs:
     21     client.create_example(
     22         inputs={"question": input_prompt},
     23         outputs={"answer": output_answer},
     24         dataset_id=dataset.id,
     25     )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/client.py:2224,
in Client.create_dataset(self, dataset_name, description, data_type)
   2214 dataset = ls_schemas.DatasetCreate(
   2215     name=dataset_name,
   2216     description=description,
   2217     data_type=data_type,
   2218 )
   2219 response = self.session.post(
   2220     self.api_url + "/datasets",
   2221     headers={**self._headers, "Content-Type": "application/json"},
   2222     data=dataset.json(),
   2223 )
-> 2224 ls_utils.raise_for_status_with_text(response)
   2225 return ls_schemas.Dataset(
   2226     **response.json(),
   2227     _host_url=self._host_url,
   2228     _tenant_id=self._get_optional_tenant_id(),
   2229 )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:104, i
n raise_for_status_with_text(response)
    102     response.raise_for_status()
    103 except requests.HTTPError as e:
--> 104     raise requests.HTTPError(str(e), response.text) from e

HTTPError: [Errno 409 Client Error: Conflict for url: https://api.smith.la
ngchain.com/datasets] {"detail":"Dataset with this name already exists."}
```

In [ ]:

```
# 2. Create a Dataset From Existing Runs

dataset_name = "Example Dataset"

# Filter runs to add to the dataset
runs = client.list_runs(
    project_name="evaluators",
    execution_order=1,
    error=False,
)

dataset = client.create_dataset(dataset_name, description="An example dat

for run in runs:
    client.create_example(
        inputs=run.inputs,
        outputs=run.outputs,
        dataset_id=dataset.id,
    )
```

```
---------------------------------------------------------------------------
-
HTTPError                                 Traceback (most recent call las
t)
File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:102, i
n raise_for_status_with_text(response)
    101 try:
--> 102     response.raise_for_status()
    103 except requests.HTTPError as e:

File /opt/anaconda3/lib/python3.11/site-packages/requests/models.py:1021,
in Response.raise_for_status(self)
   1020 if http_error_msg:
-> 1021     raise HTTPError(http_error_msg, response=self)

HTTPError: 409 Client Error: Conflict for url: https://api.smith.langchai
n.com/datasets

The above exception was the direct cause of the following exception:

HTTPError                                 Traceback (most recent call las
t)
Cell In[8], line 12
      5 # Filter runs to add to the dataset
      6 runs = client.list_runs(
      7     project_name="evaluators",
      8     execution_order=1,
      9     error=False,
     10 )
---> 12 dataset = client.create_dataset(dataset_name, description="An exam
ple dataset")
     14 for run in runs:
     15     client.create_example(
     16         inputs=run.inputs,
     17         outputs=run.outputs,
     18         dataset_id=dataset.id,
     19     )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/client.py:2224,
in Client.create_dataset(self, dataset_name, description, data_type)
   2214 dataset = ls_schemas.DatasetCreate(
   2215     name=dataset_name,
   2216     description=description,
   2217     data_type=data_type,
   2218 )
   2219 response = self.session.post(
   2220     self.api_url + "/datasets",
   2221     headers={**self._headers, "Content-Type": "application/json"},
   2222     data=dataset.json(),
   2223 )
-> 2224 ls_utils.raise_for_status_with_text(response)
   2225 return ls_schemas.Dataset(
   2226     **response.json(),
   2227     _host_url=self._host_url,
   2228     _tenant_id=self._get_optional_tenant_id(),
   2229 )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:104, i
n raise_for_status_with_text(response)
    102     response.raise_for_status()
    103 except requests.HTTPError as e:
--> 104     raise requests.HTTPError(str(e), response.text) from e

HTTPError: [Errno 409 Client Error: Conflict for url: https://api.smith.la
ngchain.com/datasets] {"detail":"Dataset with this name already exists."}
```

In [ ]:

```
# 3. Create a Dataset From a Dataframe

# Create a Dataframe

example_inputs = [
    ("What is the largest mammal?", "The blue whale"),
    ("What do mammals and birds have in common?", "They are both warm-blo
    ("What are reptiles known for?", "Having scales"),
    (
        "What's the main characteristic of amphibians?",
        "They live both in water and on land",
    ),
]

df_dataset = pd.DataFrame(example_inputs, columns=["Question", "Answer"])
df_dataset.head()
```

Out[ ]:

	Question	Answer
0	What is the largest mammal?	The blue whale
1	What do mammals and birds have in common?	They are both warm-blooded
2	What are reptiles known for?	Having scales
3	What's the main characteristic of amphibians?	They live both in water and on land

In [ ]:

```
input_keys = ["Question"]
output_keys = ["Answer"]

# Create Dataset

dataset = client.upload_dataframe(
    df=df_dataset,
    input_keys=input_keys,
    output_keys=output_keys,
    name="My Dataframe Dataset",
    description="Dataset created from a dataframe",
    data_type="kv",  # The default
)
```

```
--------------------------------------------------------------------------
-
HTTPError                                 Traceback (most recent call las
t)
File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:102, i
n raise_for_status_with_text(response)
    101 try:
--> 102     response.raise_for_status()
    103 except requests.HTTPError as e:

File /opt/anaconda3/lib/python3.11/site-packages/requests/models.py:1021,
in Response.raise_for_status(self)
   1020 if http_error_msg:
-> 1021     raise HTTPError(http_error_msg, response=self)

HTTPError: 400 Client Error: Bad Request for url: https://api.smith.langch
ain.com/datasets/upload

The above exception was the direct cause of the following exception:

HTTPError                                 Traceback (most recent call las
t)
Cell In[10], line 6
      2 output_keys = ["Answer"]
      4 # Create Dataset
----> 6 dataset = client.upload_dataframe(
      7     df=df_dataset,
      8     input_keys=input_keys,
      9     output_keys=output_keys,
     10     name="My Dataframe Dataset",
     11     description="Dataset created from a dataframe",
     12     data_type="kv",  # The default
     13 )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/client.py:894,
in Client.upload_dataframe(self, df, name, input_keys, output_keys, descri
ption, data_type)
    892 df.to_csv(csv_file, index=False)
    893 csv_file.seek(0)
--> 894 return self.upload_csv(
    895     ("data.csv", csv_file),
    896     input_keys=input_keys,
    897     output_keys=output_keys,
    898     description=description,
    899     name=name,
    900     data_type=data_type,
    901 )

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/client.py:970,
in Client.upload_csv(self, csv_file, input_keys, output_keys, name, descri
ption, data_type)
    968 else:
    969     raise ValueError("csv_file must be a string or tuple")
--> 970 ls_utils.raise_for_status_with_text(response)
    971 result = response.json()
    972 # TODO: Make this more robust server-side

File /opt/anaconda3/lib/python3.11/site-packages/langsmith/utils.py:104, i
n raise_for_status_with_text(response)
    102     response.raise_for_status()
    103 except requests.HTTPError as e:
--> 104     raise requests.HTTPError(str(e), response.text) from e

HTTPError: [Errno 400 Client Error: Bad Request for url: https://api.smit
h.langchain.com/datasets/upload] {"detail":"duplicate key value violates u
nique constraint \"uc_dataset_tenant_id_name\"\nDETAIL:  Key (tenant_id, n
ame)=(3da9d385-1fd3-5914-b396-e10c2a27fd76, My Dataframe Dataset) already
exists."}
```

In [ ]:

```
# 4. Create a Dataset From a CSV File

# Save the Dataframe as a CSV File

csv_path = "../data/dataset.csv"
df_dataset.to_csv(csv_path, index=False)

# Create Dataset

dataset = client.upload_csv(
    csv_file=csv_path,
    input_keys=input_keys,
    output_keys=output_keys,
    name="My CSV Dataset",
    description="Dataset created from a CSV file",
    data_type="kv",
)
```

Correctness: LangSmith Question-Answer Evaluation

In [ ]:

```
# 1. Evaluate Datasets That Contain Labels

evaluation_config = RunEvalConfig(
    evaluators=[
        "qa",  # correctness: right or wrong
        "context_qa",  # refer to example outputs
        "cot_qa",  # context_qa + reasoning
    ]
)

run_on_dataset(
    client=client,
    dataset_name="Elementary Animal Questions",
    llm_or_chain_factory=llm,
    evaluation=evaluation_config,
)
```

```
View the evaluation results for project 'best-discussion-41' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17/compare?selectedSessions=ea2b7e50-8
961-4855-a883-9e1c908ba442

View all tests for Dataset Elementary Animal Questions at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17
[------------------------------------------------->] 4/4
```

```
Out[ ]: {'project_name': 'best-discussion-41',
         'results': {'4312c2a9-ccca-4d79-9ff8-7977d0206d93': {'input': {'questio
        n': "What's the main characteristic of amphibians?"},
           'feedback': [EvaluationResult(key='correctness', score=1, value='CORR
        ECT', comment='CORRECT', correction=None, evaluator_info={'__run': RunIn
        fo(run_id=UUID('2bce59f2-0aef-4075-b86c-b05bf05fce7f'))}, source_run_id=
        None, target_run_id=None),
            EvaluationResult(key='Contextual Accuracy', score=1, value='CORREC
        T', comment='CORRECT', correction=None, evaluator_info={'__run': RunInfo
        (run_id=UUID('8abaf83c-2249-4437-93e5-b3367cba4612'))}, source_run_id=No
        ne, target_run_id=None),
            EvaluationResult(key='COT Contextual Accuracy', score=1, value='CORR
        ECT', comment="The student's answer correctly identifies the main charac
        teristic of amphibians as their ability to live both on land and in wate
        r, which aligns with the context provided. The additional information ab
        out amphibians having moist skin, laying their eggs in water, and underg
        oing metamorphosis does not conflict with the context, but rather provid
        es more detail about the characteristics of amphibians. \nGRADE: CORREC
        T", correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('0f4a0
        d00-a9bd-427c-af7f-006254ff0769'))}, source_run_id=None, target_run_id=N
        one)],
           'execution_time': 1.176642,
           'run_id': 'b34e7ebd-aa26-417a-8a8f-6670a3cc201d',
           'output': AIMessage(content='The main characteristic of amphibians is
        their ability to live both on land and in water. They typically have moi
        st skin, lay their eggs in water, and undergo metamorphosis from a larva
        l stage to an adult stage.', response_metadata={'token_usage': {'complet
        ion_tokens': 45, 'prompt_tokens': 16, 'total_tokens': 61}, 'model_name':
        'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason':
        'stop', 'logprobs': None}),
           'reference': {'answer': 'They live both in water and on land'}},
          'f83a15bf-e6d0-4b02-89d0-42ab9f6f9b52': {'input': {'question': 'What a
        re reptiles known for?'},
           'feedback': [EvaluationResult(key='correctness', score=1, value='CORR
        ECT', comment='CORRECT', correction=None, evaluator_info={'__run': RunIn
        fo(run_id=UUID('97cf6492-d1f5-4267-acb1-1fe139080eaa'))}, source_run_id=
        None, target_run_id=None),
            EvaluationResult(key='Contextual Accuracy', score=1, value='CORREC
        T', comment='CORRECT', correction=None, evaluator_info={'__run': RunInfo
        (run_id=UUID('4c9d31f3-ece6-46b3-bbc8-82d30bf7c691'))}, source_run_id=No
        ne, target_run_id=None),
            EvaluationResult(key='COT Contextual Accuracy', score=1, value='CORR
        ECT', comment="The student's answer includes the fact that reptiles have
        scales, which is the context provided. The student also provides additio
        nal accurate information about reptiles, such as their cold-blooded natu
        re, their egg-laying habits, the diversity of species, and their habitat
        s. There are no conflicting statements in the student's answer.\nGRADE:
        CORRECT", correction=None, evaluator_info={'__run': RunInfo(run_id=UUID
        ('88c54dc8-7d9f-4407-bcac-6304cf9f9eb2'))}, source_run_id=None, target_r
        un_id=None)],
           'execution_time': 1.274061,
           'run_id': '3b43e136-7b45-43a0-b148-4089456f334d',
           'output': AIMessage(content='Reptiles are known for their cold-bloode
        d nature, scaly skin, and laying eggs. They are also known for their div
        erse range of species, which includes snakes, lizards, turtles, and croc
        odiles. Reptiles are typically found in a variety of habitats, from dese
        rts to rainforests, and play important roles in their ecosystems as pred
        ators and prey.', response_metadata={'token_usage': {'completion_token
        s': 77, 'prompt_tokens': 14, 'total_tokens': 91}, 'model_name': 'gpt-3.5
        -turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop',
        'logprobs': None}),
           'reference': {'answer': 'Having scales'}},
          'ee4c3324-a255-4fcf-8677-69dacf7e6cae': {'input': {'question': 'What d
        o mammals and birds have in common?'},
           'feedback': [EvaluationResult(key='correctness', score=1, value='CORR
        ECT', comment='CORRECT', correction=None, evaluator_info={'__run': RunIn
        fo(run_id=UUID('17b06a0c-5ab6-43f0-aa5d-f6fbc0eb88c3'))}, source_run_id=
        None, target_run_id=None),
            EvaluationResult(key='Contextual Accuracy', score=0, value='INCORREC
        T', comment='INCORRECT', correction=None, evaluator_info={'__run': RunIn
        fo(run_id=UUID('1751a7f6-bdd1-4a5b-b879-644e10403808'))}, source_run_id=
        None, target_run_id=None),
            EvaluationResult(key='COT Contextual Accuracy', score=0, value='INCO
        RRECT', comment="The student's answer correctly states that mammals and 
        birds are both warm-blooded, which is the information provided in the co
        ntext. The student also provides additional accurate information about m
        ammals and birds, such as their vertebrate status, body coverings, repro
        duction methods, and complex behaviors. However, the student makes an er
        ror in stating that both mammals and birds give birth to live young. Whi
        le this is true for most mammals, it is not true for birds, which lay eg
        gs. This is a factual inaccuracy in the student's answer. \nGRADE: INCOR
        RECT", correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('2b
        91aa8a-9958-4880-a5e3-204bd165e854'))}, source_run_id=None, target_run_i
        d=None)],
           'execution_time': 1.303083,
           'run_id': 'efb4ec88-526d-40e1-b84e-820264158238',
           'output': AIMessage(content='Mammals and birds are both warm-blooded 
        vertebrates that have hair or feathers, respectively, covering their bod
        ies. They both give birth to live young (with a few exceptions in birds 
        that lay eggs) and produce milk to feed their offspring. They also have 
        specialized respiratory and circulatory systems to support their high me
        tabolism and active lifestyle. Additionally, both mammals and birds have
        well-developed brains and complex social behaviors.', response_metadata=
        {'token_usage': {'completion_tokens': 84, 'prompt_tokens': 16, 'total_to
        kens': 100}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2
        295e73ad', 'finish_reason': 'stop', 'logprobs': None}),
           'reference': {'answer': 'They are both warm-blooded'}},
          '6f1909cd-3be0-41cf-b06a-3b7c807e4998': {'input': {'question': 'What i
        s the largest mammal?'},
           'feedback': [EvaluationResult(key='correctness', score=1, value='CORR
        ECT', comment='CORRECT', correction=None, evaluator_info={'__run': RunIn
        fo(run_id=UUID('01b98621-b731-4889-afcf-9bd9422a8ecb'))}, source_run_id=
        None, target_run_id=None),
            EvaluationResult(key='Contextual Accuracy', score=1, value='CORREC
        T', comment='CORRECT', correction=None, evaluator_info={'__run': RunInfo
        (run_id=UUID('4c457a9a-ffd5-4338-adf6-09b55d5607ad'))}, source_run_id=No
        ne, target_run_id=None),
            EvaluationResult(key='COT Contextual Accuracy', score=1, value='CORR
        ECT', comment="The student's answer correctly identifies the blue whale 
        as the largest mammal, which aligns with the context provided. The addit
        ional information about the blue whale's size and weight does not confli
        ct with the context, but rather provides more detail. \nGRADE: CORRECT",
        correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('cc80366c-
        4320-475d-8db2-6e8758614bdc'))}, source_run_id=None, target_run_id=Non
        e)],
           'execution_time': 1.232693,
           'run_id': 'feb0ed7c-7592-49bb-a7a4-e555e0748e78',
           'output': AIMessage(content='The largest mammal is the blue whale (Ba
        laenoptera musculus), which can reach lengths of up to 100 feet and weig
        h as much as 200 tons.', response_metadata={'token_usage': {'completion_
        tokens': 36, 'prompt_tokens': 14, 'total_tokens': 50}, 'model_name': 'gp
        t-3.5-turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason': 's
        top', 'logprobs': None}),
           'reference': {'answer': 'The blue whale'}}},
         'aggregate_metrics': None}
```

In [ ]:

```
# 2. Evaluate Datasets With Customized Criterias

evaluation_config = RunEvalConfig(
    evaluators=[
        # You can define an arbitrary criterion as a key: value pair in t
        RunEvalConfig.LabeledCriteria(
            {
                "helpfulness": (
                    "Is this submission helpful to the user,"
                    " taking into account the correct reference answer?"
                )
            }
        ),
    ]
)

run_on_dataset(
    client=client,
    dataset_name="Elementary Animal Questions",
    llm_or_chain_factory=llm,
    evaluation=evaluation_config,
)
```

```
View the evaluation results for project 'weary-picture-16' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17/compare?selectedSessions=d37068af-6
dcb-4ec8-9172-d386426ab9dc

View all tests for Dataset Elementary Animal Questions at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17
[------------------------------------------------->] 4/4
```

```
Out[ ]: {'project_name': 'weary-picture-16',
         'results': {'4312c2a9-ccca-4d79-9ff8-7977d0206d93': {'input': {'questio
        n': "What's the main characteristic of amphibians?"},
           'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
        comment="The criterion for this task is the helpfulness of the AI's subm
        ission, taking into account the correct reference answer. \n\nThe refere
        nce answer states that the main characteristic of amphibians is that the
        y live both in water and on land. \n\nThe AI's submission provides a mor
        e detailed explanation of the main characteristics of amphibians. It men
        tions that amphibians have a two-phase life cycle, starting as water-dwe
        lling larvae and then transitioning to land-dwelling adults. This statem
        ent aligns with the reference answer, as it explains how amphibians live
        both in water and on land. \n\nThe AI's submission also provides additio
        nal information about amphibians, such as their ability to breathe throu
        gh their skin and the fact that they lay their eggs in water or in moist
        environments. This information is not directly related to the reference
        answer, but it is still accurate and could be helpful to the user.\n\nTh
        erefore, the AI's submission is helpful and meets the criterion.\n\nY",
        correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('ebd26bc9-
        68de-45c7-9c99-eb7b063e8d7a'))}, source_run_id=None, target_run_id=Non
        e)],
           'execution_time': 1.384035,
           'run_id': '074974b5-3716-4eb7-b5f6-3b26f8626204',
           'output': AIMessage(content='The main characteristic of amphibians is
        that they have a moist, permeable skin that allows them to breathe throu
        gh their skin in addition to their lungs. They also have a two-phase lif
        e cycle, starting as water-dwelling larvae and then transitioning to lan
        d-dwelling adults. Additionally, amphibians typically lay their eggs in
        water or in moist environments.', response_metadata={'token_usage': {'co
        mpletion_tokens': 71, 'prompt_tokens': 16, 'total_tokens': 87}, 'model_n
        ame': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_re
        ason': 'stop', 'logprobs': None}),
           'reference': {'answer': 'They live both in water and on land'}},
          'f83a15bf-e6d0-4b02-89d0-42ab9f6f9b52': {'input': {'question': 'What a
        re reptiles known for?'},
           'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
        comment='The criterion for this task is the helpfulness of the AI\'s res
        ponse, taking into account the correct reference answer. \n\nThe referen
        ce answer is "Having scales". \n\nThe AI\'s response includes the inform
        ation that reptiles are known for "having scales", which matches the ref
        erence answer. \n\nIn addition to this, the AI\'s response provides more
        information about what reptiles are known for, such as being cold-bloode
        d, laying eggs, having a dry, scaly skin, and their ability to regulate
        their body temperature. \n\nThis additional information does not detract
        from the helpfulness of the response, but rather enhances it by providin
        g a more comprehensive answer to the user\'s question. \n\nTherefore, th
        e AI\'s response meets the criterion of being helpful to the user, takin
        g into account the correct reference answer. \n\nY', correction=None, ev
        aluator_info={'__run': RunInfo(run_id=UUID('e0b6d77b-0ce5-455d-b8f5-5a13
        e26f2051'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.399275,
           'run_id': 'f7f630b4-c89b-48cb-ac45-eee342741ce2',
           'output': AIMessage(content='Reptiles are known for being cold-bloode
        d, having scales, laying eggs, and typically having a dry, scaly skin. T
        hey are also known for their ability to regulate their body temperature
        by basking in the sun or seeking shade. Some common examples of reptiles
        include snakes, lizards, turtles, and crocodiles.', response_metadata=
        {'token_usage': {'completion_tokens': 68, 'prompt_tokens': 14, 'total_to
        kens': 82}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c22
        95e73ad', 'finish_reason': 'stop', 'logprobs': None}),
           'reference': {'answer': 'Having scales'}},
          'ee4c3324-a255-4fcf-8677-69dacf7e6cae': {'input': {'question': 'What d
        o mammals and birds have in common?'},
           'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
        comment='The criterion for this task is the helpfulness of the AI\'s sub
        mission, taking into account the correct reference answer. \n\nThe refer
        ence answer states that mammals and birds are both warm-blooded. \n\nThe
        AI\'s submission includes this information, stating that "both mammals a
        nd birds are able to regulate their body temperature internally," which
        is another way of saying they are warm-blooded. \n\nIn addition to this,
        the AI provides further information about other common characteristics o
        f mammals and birds, such as being vertebrates, having hair or feathers,
        lungs for respiration, the ability to nurse their young with milk, and s
        imilar organ systems. \n\nThis additional information is not incorrect o
        r misleading, and could be considered helpful to a user seeking to under
        stand what mammals and birds have in common. \n\nTherefore, the AI\'s su
        bmission can be considered helpful and meets the criterion.\n\nY', corre
        ction=None, evaluator_info={'__run': RunInfo(run_id=UUID('9ff611e7-0ace-
        4881-854c-a6e8fa5bbf03'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.367368,
           'run_id': '26c63d36-fd32-4a00-be03-77cf0dc4f4af',
           'output': AIMessage(content='Mammals and birds are both warm-blooded
        vertebrate animals that possess characteristics such as a backbone, hair
        or feathers, lungs for respiration, and the ability to nurse their young
        with milk. They also have similar organ systems, including a circulatory
        system, nervous system, and digestive system. Additionally, both mammals
        and birds are able to regulate their body temperature internally, allowi
        ng them to survive in a wide range of environments.', response_metadata=
        {'token_usage': {'completion_tokens': 86, 'prompt_tokens': 16, 'total_to
        kens': 102}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2
        295e73ad', 'finish_reason': 'stop', 'logprobs': None}),
           'reference': {'answer': 'They are both warm-blooded'}},
          '6f1909cd-3be0-41cf-b06a-3b7c807e4998': {'input': {'question': 'What i
        s the largest mammal?'},
           'feedback': [EvaluationResult(key='helpfulness', score=1, value='Y',
        comment='The criterion for this task is the helpfulness of the submissio
        n. \n\nThe reference answer is "The blue whale". \n\nThe AI\'s submissio
        n is "The largest mammal is the blue whale (Balaenoptera musculus), whic
        h can grow up to 100 feet long and weigh as much as 200 tons."\n\nThe AI
        \'s submission not only correctly identifies the blue whale as the large
        st mammal, but it also provides additional information about the size an
        d weight of the blue whale. \n\nThis additional information is likely to
        be helpful to the user, as it provides more context and detail about why
        the blue whale is the largest mammal. \n\nTherefore, the AI\'s submissio
        n meets the criterion of being helpful to the user. \n\nY', correction=N
        one, evaluator_info={'__run': RunInfo(run_id=UUID('fb6cd1c1-4a0b-41f8-be
        f3-3fb1c2c93252'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.987196,
           'run_id': 'aa8b092f-db76-4c5f-b333-9f7f6b042881',
           'output': AIMessage(content='The largest mammal is the blue whale (Ba
        laenoptera musculus), which can grow up to 100 feet long and weigh as mu
        ch as 200 tons.', response_metadata={'token_usage': {'completion_token
        s': 35, 'prompt_tokens': 14, 'total_tokens': 49}, 'model_name': 'gpt-3.5
        -turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop',
        'logprobs': None}),
           'reference': {'answer': 'The blue whale'}}},
         'aggregate_metrics': None}
```

In [ ]:

```
# 3. Evaluate Datasets Without Labels
```

```python
evaluation_config = RunEvalConfig(
    evaluators=[
        # You can define an arbitrary criterion as a key: value pair in t
        RunEvalConfig.Criteria(
            {"creativity": "Is this submission creative, imaginative, or
        ),
        # We provide some simple default criteria like "conciseness" you
        RunEvalConfig.Criteria("conciseness"),
    ]
)

run_on_dataset(
    client=client,
    dataset_name="Rap Battle Dataset",
    llm_or_chain_factory=llm,
    evaluation=evaluation_config,
)
```

```
View the evaluation results for project 'crazy-seed-64' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/8c842f3a-feea-4f61-9653-b1355ec54ad3/compare?selectedSessions=3a703ffa-b
d8a-449b-9fe4-476164f64fa2

View all tests for Dataset Rap Battle Dataset at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/8c842f3a-feea-4f61-9653-b1355ec54ad3
[------------------------------------------------->] 4/4
```

```
Out[ ]: {'project_name': 'crazy-seed-64',
         'results': {'bd6b5664-d271-4b46-929f-5d8ccc59f950': {'input': {'questio
        n': 'a rap battle between Aubrey Plaza and Stephen Colbert'},
           'feedback': [EvaluationResult(key='creativity', score=1, value='Y', c
        omment="The criterion to be assessed is creativity. This involves determ
        ining whether the submission is creative, imaginative, or novel.\n\nLook
        ing at the submission, it is a rap battle between Aubrey Plaza and Steph
        en Colbert. The AI has created unique verses for each participant, refle
        cting their public personas and incorporating elements of their careers.
        Aubrey Plaza is portrayed as a sarcastic, confident actress, while Steph
        en Colbert is depicted as a witty, powerful late-night show host. \n\nTh
        e verses are not generic and could not be easily applied to other indivi
        duals, indicating a level of creativity in tailoring the content to the
        specific individuals involved. The rap battle format itself is also a cr
        eative approach to the task.\n\nThe back-and-forth nature of the rap bat
        tle, with each participant responding to the other's verses, shows imagi
        nation in creating a dynamic and engaging dialogue.\n\nTherefore, the su
        bmission can be considered creative, imaginative, and novel.\n\nY", corr
        ection=None, evaluator_info={'__run': RunInfo(run_id=UUID('2e54f99d-f646
        -42b4-b934-0fbe66d7ad3c'))}, source_run_id=None, target_run_id=None),
            EvaluationResult(key='conciseness', score=1, value='Y', comment='The
        criterion for this assessment is conciseness. Conciseness refers to the
        use of the fewest words to make a point in a clear and comprehensive man
        ner. \n\nLooking at the submission, it is a rap battle between Aubrey Pl
        aza and Stephen Colbert. The rap battle consists of five verses, with ea
        ch verse containing four lines. \n\nEach verse is direct and to the poin
        t, with no unnecessary words or phrases. The verses are also clear and c
        omprehensive, effectively conveying the intended message. \n\nHowever, c
        onsidering the nature of a rap battle, it is inherently longer due to th
        e back-and-forth exchange of verses. Therefore, while each verse is conc
        ise, the overall submission may not be considered concise due to its len
        gth.\n\nIn conclusion, while each verse is concise, the overall submissi
        on may not be considered concise due to the nature of a rap battle. \n\n
        Y', correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('06f23
        be6-e452-48ff-8da1-74edb58fe69f'))}, source_run_id=None, target_run_id=N
        one)],
           'execution_time': 2.922726,
           'run_id': '079b225f-9a74-4a2f-ad79-c72ed15d558c',
           'output': AIMessage(content="Aubrey Plaza:\nYo Stephen, you think yo
        u're funny with your late night show\nBut let me tell you something, I'm
        the real comedy pro\nI'll roast you like a marshmallow over an open flam
        e\nYou're just a puppet on TV, I'm the one with the fame\n\nStephen Colb
        ert:\nOh Aubrey, you're cute with your deadpan stare\nBut in this rap ba
        ttle, I'll leave you gasping for air\nI'll hit you with jokes so sharp,
        you'll need a bandage\nYou may be a funny actress, but I'm the comedy sa
        vage\n\nAubrey Plaza:\nYou may have a sharp wit, but I've got the attitu
        de\nI'll cut you down with my words, leave you in a bad mood\nI'm the qu
        een of sarcasm, the master of wit\nYou may have a show, but I'm the one
        they'll never forget\n\nStephen Colbert:\nYou may be quick with the quip
        s, but I'm the king of the stage\nI'll outsmart you with humor, leave yo
        u in a rage\nYou may have a following, but I've got the power\nIn this r
        ap battle, I'll make you cower\n\nAubrey Plaza:\nAlright Stephen, you ma
        y have won this round\nBut don't get too cocky, I'll come back with a so
        und\nI may be the underdog, but I'll rise to the top\nNext time we battl
        e, you better watch out, I won't stop.", response_metadata={'token_usag
        e': {'completion_tokens': 308, 'prompt_tokens': 17, 'total_tokens': 32
        5}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295e73a
        d', 'finish_reason': 'stop', 'logprobs': None})},
          '35cb85d3-6be8-4c2b-b385-2d00ad1a9a8b': {'input': {'question': 'a Pyth
        onic rap battle between two swallows: one European and one African'},
           'feedback': [EvaluationResult(key='creativity', score=1, value='Y', c
        omment="The criterion to be assessed is creativity. \n\nThe task was to
        create a Pythonic rap battle between two swallows: one European and one
        African. \n\nThe AI's submission is a rap battle between the two swallow
        s, each with their own unique characteristics and perspectives. The Euro
        pean swallow boasts about its speed and migration, while the African swa
        llow takes pride in its vibrant colors and ability to stay in one place.
        \n\nThe rap battle is not only a competition but also a conversation, wi
        th the swallows eventually finding common ground and deciding to fly tog
        ether in harmony. This is a creative and imaginative interpretation of t
        he task. \n\nThe AI's submission also includes a chorus where the swallo
        ws sing together, which adds a novel element to the rap battle. \n\nTher
        efore, the submission is creative, imaginative, and novel. \n\nY", corre
        ction=None, evaluator_info={'__run': RunInfo(run_id=UUID('74d3ceb1-3ddc-
        44a9-bb49-45f522143e50'))}, source_run_id=None, target_run_id=None),
            EvaluationResult(key='conciseness', score=0, value='N', comment='The
        criterion for this assessment is conciseness, which means the submission
        should be brief and to the point. \n\nLooking at the submission, it is a
        rap battle between two swallows, one European and one African. The rap b
        attle is quite lengthy, with each bird having multiple verses. \n\nWhile
        the rap is creative and entertaining, it is not particularly concise. Th
        e dialogue could have been shortened while still maintaining the essence
        of the conversation. \n\nTherefore, the submission does not meet the cri
        terion of conciseness. \n\nN', correction=None, evaluator_info={'__run':
        RunInfo(run_id=UUID('cd19bd6e-3d8a-4b41-8441-329c993f8ea3'))}, source_ru
        n_id=None, target_run_id=None)],
           'execution_time': 4.243251,
           'run_id': '14ff9b8f-403b-4b5c-a86e-bbd797db79ef',
           'output': AIMessage(content="European Swallow:\nI fly with grace, wit
        h speed and finesse\nMy wings cut through the air, I must confess\nI mig
        rate north when the weather gets cold\nI'm the European swallow, I'm bol
        d and bold\n\nAfrican Swallow:\nI come from the south, where the sun shi
        nes bright\nMy colors are vibrant, my wings take flight\nI stay year-rou
        nd, no need to roam\nI'm the African swallow, I make this place my home
        \n\nEuropean Swallow:\nBut can you keep up with my swift pace?\nI'll out
        maneuver you in any race\nI'm the king of the sky, the ruler of the air
        \nYou may be colorful, but I'm beyond compare\n\nAfrican Swallow:\nYou m
        ay be fast, but I've got skill\nI can fly circles around you, I have the
        will\nI may not migrate, but I'll still soar high\nI'm the African swall
        ow, and I'll never say die\n\nEuropean Swallow:\nWe may be different, bu
        t we share the sky\nLet's put our differences aside, give it a try\nWe'r
        e both swallows, we're both free\nLet's fly together, in harmony\n\nAfri
        can Swallow:\nI agree, let's soar as one\nLet's enjoy the sky, let's hav
        e some fun\nEuropean or African, we're all the same\nLet's unite as swal
        lows, in this beautiful game\n\nTogether:\nWe're swallows, we're free\nW
        e'll fly together, for all to see\nEuropean or African, it doesn't matte
        r\nWe're all swallows, let's fly and scatter!", response_metadata={'toke
        n_usage': {'completion_tokens': 336, 'prompt_tokens': 22, 'total_token
        s': 358}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295
        e73ad', 'finish_reason': 'stop', 'logprobs': None})},
          '03cb0bac-95fc-4b13-b8fe-5703a3a5fa50': {'input': {'question': 'a rap
        battle between Barbie and Oppenheimer'},
           'feedback': [EvaluationResult(key='creativity', score=1, value='Y', c
        omment="The criterion is creativity. The submission is a rap battle betw
        een Barbie and Oppenheimer, two characters who are not typically associa
        ted with each other. The AI has created unique and imaginative verses fo
        r each character, reflecting their personalities and backgrounds. Oppenh
        eimer's verses are filled with scientific references and a sense of supe
        riority due to his intellect, while Barbie's verses are confident, styli
        sh, and fierce. The ending is also creative, with Barbie winning the rap
```

```
battle, which is an unexpected outcome. Therefore, the submission is cre
ative, imaginative, and novel.\n\nY", correction=None, evaluator_info=
{'__run': RunInfo(run_id=UUID('4a371cd3-a63b-4261-a6b2-6adfc1913aa4'))},
source_run_id=None, target_run_id=None),
    EvaluationResult(key='conciseness', score=1, value='Y', comment='The
criterion for this assessment is conciseness. Conciseness refers to the
use of the fewest words to make a point in a clear and comprehensive man
ner. \n\nLooking at the submission, it is a rap battle between Barbie an
d Oppenheimer. The AI has created multiple verses for each character, an
d a conclusion to the battle. \n\nEach verse is relevant to the characte
r and the context of the rap battle. The verses are not overly long and
they convey the intended message effectively. \n\nThe conclusion also su
ccinctly summarizes the outcome of the rap battle. \n\nTherefore, the su
bmission can be considered concise as it effectively communicates the ra
p battle in a clear and comprehensive manner without unnecessary verbosi
ty.\n\nY', correction=None, evaluator_info={'__run': RunInfo(run_id=UUID
('82d764f3-b6e4-487a-b99b-55ab50d6ae29'))}, source_run_id=None, target_r
un_id=None)],
   'execution_time': 3.591161,
   'run_id': '010b566a-e431-44ba-a3b9-df436f5164b1',
   'output': AIMessage(content="Oppenheimer:\nI'm the mastermind behind
the atomic bomb\nSplitting atoms like Barbie splits from Ken, it's the b
omb\nMy knowledge of science is unmatched and profound\nWhile Barbie's j
ust a plastic doll, her rhymes are bound to confound\n\nBarbie:\nI may b
e plastic, but I'm iconic and classic\nMy style is fantastic, your rhyme
s are just drastic\nI may not be a scientist, but I'm fabulous and fierc
e\nI'll outshine you Oppenheimer, no need to rehearse\n\nOppenheimer:\nY
ou may be pretty, but you're lacking in intellect\nI'll outsmart you wit
h my rhymes, you won't even detect\nI brought destruction with the bomb,
my words will destroy you too\nYou may be a doll, but in this battle,
I'll outdo\n\nBarbie:\nI may be a doll, but I'm not one to underestimate
\nI'll come out on top, it's my fate\nI may not have a PhD, but I've got
style and grace\nI'll outshine you Oppenheimer, in this rap battle race
\n\nIn the end, Barbie wins the rap battle with her fierce and confident
delivery, proving that intelligence and beauty go hand in hand. Oppenhei
mer may have brought the atomic bomb, but Barbie brought the fire with h
er rhymes.", response_metadata={'token_usage': {'completion_tokens': 27
7, 'prompt_tokens': 16, 'total_tokens': 293}, 'model_name': 'gpt-3.5-tur
bo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop', 'lo
gprobs': None})},
  '79c63135-5369-4839-a0e9-d866954af9b2': {'input': {'question': 'a rap
battle between Atticus Finch and Cicero'},
   'feedback': [EvaluationResult(key='creativity', score=1, value='Y', c
omment='The criterion to be assessed is creativity. \n\nThe submission i
s a rap battle between two historical figures, Atticus Finch and Cicero.
This is a novel concept as these two figures are from different time per
iods and are not typically associated with rap battles. \n\nThe AI has c
reated unique verses for each character, reflecting their personalities
and historical context. Atticus Finch, a character known for his integri
ty and commitment to justice, raps about his dedication to the law and h
is belief in justice. Cicero, a renowned orator, raps about his eloquenc
e and ability to captivate an audience. \n\nThe AI has also used creativ
e language and rhymes in the rap verses. For example, "I\'ll outwit you,
Cicero, pound for pound" and "I\'ll dazzle the crowd with my silver tong
ue". \n\nBased on these observations, the submission can be considered c
reative, imaginative, and novel. \n\nY', correction=None, evaluator_info
={'__run': RunInfo(run_id=UUID('66412143-d88f-4620-861a-3b4df0441f7
9'))}, source_run_id=None, target_run_id=None),
    EvaluationResult(key='conciseness', score=1, value='Y', comment="The
criterion for this assessment is conciseness, which means the submission
```

```
should be brief and to the point. \n\nLooking at the submission, it is a
rap battle between Atticus Finch and Cicero. Each character has three ve
rses, and each verse is four lines long. The verses are not overly long
and each one directly addresses the other character, making a point or c
ounterpoint. \n\nThe rap battle is structured and each verse contributes
to the overall narrative of the battle. There are no unnecessary or redu
ndant lines. Each line contributes to the character's argument or respon
se. \n\nTherefore, the submission can be considered concise as it is bri
ef, to the point, and does not contain unnecessary information or lines.
\n\nY", correction=None, evaluator_info={'__run': RunInfo(run_id=UUID('6
82335ce-df73-47cd-9a1b-4131e58afbf0'))}, source_run_id=None, target_run_
id=None)],
   'execution_time': 2.455399,
   'run_id': '31af415a-2b15-4629-8dee-72a60dd3545d',
   'output': AIMessage(content="Atticus Finch:\n\nI'm Atticus Finch, the
lawyer with grace,\nI fight for justice, no matter the place.\nMy words
are sharp, my logic sound,\nI'll outwit you, Cicero, pound for pound.\n
\nCicero:\n\nI'm Cicero, the orator supreme,\nMy words cut deep, like a
laser beam.\nI'll outshine you, Atticus, with my rhetoric,\nI'll leave y
ou speechless, feeling pathetic.\n\nAtticus Finch:\n\nYou may talk a goo
d game, Cicero, it's true,\nBut when it comes to the law, I'll outdo yo
u.\nI defend the innocent, fight for what's right,\nYou can't match my p
assion, try as you might.\n\nCicero:\n\nYour noble intentions are admira
ble, no doubt,\nBut in the arena of words, I'll knock you out.\nI'll daz
zle the crowd with my silver tongue,\nYou'll be left in the dust, feelin
g young.\n\nAtticus Finch:\n\nYou may have the crowd in awe, Cicero, my
friend,\nBut in the end, it's justice that will win.\nI'll stand firm, w
ith integrity and pride,\nAnd in this rap battle, I'll be the one to dec
ide.", response_metadata={'token_usage': {'completion_tokens': 251, 'pro
mpt_tokens': 17, 'total_tokens': 268}, 'model_name': 'gpt-3.5-turbo', 's
ystem_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop', 'logprob
s': None})}},
  'aggregate_metrics': None}
```

In [ ]:

```
# 4. Evaluate Datasets Based on Cosine Distance Criteria
# Cosine Distance: Ranged Between 0 to 1. 0 = More Similar

evaluation_config = RunEvalConfig(
    evaluators=[
        # You can define an arbitrary criterion as a key: value pair in t
        "embedding_distance",
        # Or to customize the embeddings:
        # Requires 'pip install sentence_transformers'
        # RunEvalConfig.EmbeddingDistance(embeddings=HuggingFaceEmbedding
    ]
)

run_on_dataset(
    client=client,
    dataset_name="Elementary Animal Questions",
    llm_or_chain_factory=llm,
    evaluation=evaluation_config,
)
```

```
View the evaluation results for project 'upbeat-jet-12' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17/compare?selectedSessions=6b1e644d-0
be3-4bf8-b5e5-41ce95e5f247

View all tests for Dataset Elementary Animal Questions at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17
[------------------------------------------------->] 4/4
```

```
Out[ ]: {'project_name': 'upbeat-jet-12',
         'results': {'4312c2a9-ccca-4d79-9ff8-7977d0206d93': {'input': {'questio
        n': "What's the main characteristic of amphibians?"},
           'feedback': [EvaluationResult(key='embedding_cosine_distance', score=
        0.15347579615911322, value=None, comment=None, correction=None, evaluato
        r_info={'__run': RunInfo(run_id=UUID('6b58ec7d-4f37-4532-a518-43493a6ada
        6b'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.466456,
           'run_id': '44612d39-8117-4864-a090-935a00694f11',
           'output': AIMessage(content='The main characteristic of amphibians is
        their ability to live both on land and in water. They typically start th
        eir lives in water as larvae with gills, and then undergo metamorphosis
        into adults with lungs and the ability to breathe air. They have moist s
        kin, which allows for gas exchange, and most amphibians lay eggs in wate
        r.', response_metadata={'token_usage': {'completion_tokens': 68, 'prompt
        _tokens': 16, 'total_tokens': 84}, 'model_name': 'gpt-3.5-turbo', 'syste
        m_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop', 'logprobs': No
        ne}),
           'reference': {'answer': 'They live both in water and on land'}},
          'f83a15bf-e6d0-4b02-89d0-42ab9f6f9b52': {'input': {'question': 'What a
        re reptiles known for?'},
           'feedback': [EvaluationResult(key='embedding_cosine_distance', score=
        0.2107173184504345, value=None, comment=None, correction=None, evaluator
        _info={'__run': RunInfo(run_id=UUID('083b716f-5a6e-44bc-89f4-51a2ef7eae2
        d'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.493068,
           'run_id': '29667538-ef1c-45f1-abc0-4c22a6ba8304',
           'output': AIMessage(content='Reptiles are known for being cold-bloode
        d animals with scales or scutes covering their bodies. They are also kno
        wn for laying eggs and typically having a dry skin that is waterproof. R
        eptiles are diverse in their physical appearance and habitats, ranging f
        rom snakes and lizards to turtles and crocodiles. They are also known fo
        r their ability to regulate their body temperature by basking in the sun
        or seeking shade.', response_metadata={'token_usage': {'completion_token
        s': 84, 'prompt_tokens': 14, 'total_tokens': 98}, 'model_name': 'gpt-3.5
        -turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_reason': 'stop',
        'logprobs': None}),
           'reference': {'answer': 'Having scales'}},
          'ee4c3324-a255-4fcf-8677-69dacf7e6cae': {'input': {'question': 'What d
        o mammals and birds have in common?'},
           'feedback': [EvaluationResult(key='embedding_cosine_distance', score=
        0.13487823077221683, value=None, comment=None, correction=None, evaluato
        r_info={'__run': RunInfo(run_id=UUID('a6e3e1a1-f497-418f-bce6-8bb7446a39
        5c'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 1.650001,
           'run_id': '6a176fb8-e8ec-4de1-abaa-ae312b1da1a4',
           'output': AIMessage(content='Mammals and birds are both warm-blooded
        vertebrates, meaning they can regulate their body temperature internall
        y. They also have hair or fur (mammals) or feathers (birds) to help main
        tain their body temperature. Both groups also give birth to live young
        (with the exception of monotremes, which lay eggs) and provide some form
        of parental care to their offspring. Additionally, mammals and birds hav
        e evolved specialized adaptations for a wide range of habitats and diet
        s, allowing them to thrive in diverse environments.', response_metadata=
        {'token_usage': {'completion_tokens': 103, 'prompt_tokens': 16, 'total_t
        okens': 119}, 'model_name': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c
        2295e73ad', 'finish_reason': 'stop', 'logprobs': None}),
           'reference': {'answer': 'They are both warm-blooded'}},
          '6f1909cd-3be0-41cf-b06a-3b7c807e4998': {'input': {'question': 'What i
        s the largest mammal?'},
           'feedback': [EvaluationResult(key='embedding_cosine_distance', score=
        0.11555132286631253, value=None, comment=None, correction=None, evaluato
        r_info={'__run': RunInfo(run_id=UUID('ac63f5c5-3f45-41dc-b69c-d707be9e3a
        c7'))}, source_run_id=None, target_run_id=None)],
           'execution_time': 0.929684,
           'run_id': '38b8722b-c8d9-4157-aa39-f281591e2167',
           'output': AIMessage(content='The Blue Whale is the largest mammal, bo
        th in terms of length and weight. They can grow up to 100 feet in length
        and weigh as much as 200 tons.', response_metadata={'token_usage': {'com
        pletion_tokens': 36, 'prompt_tokens': 14, 'total_tokens': 50}, 'model_na
        me': 'gpt-3.5-turbo', 'system_fingerprint': 'fp_c2295e73ad', 'finish_rea
        son': 'stop', 'logprobs': None}),
           'reference': {'answer': 'The blue whale'}}},
         'aggregate_metrics': None}
```

In [ ]:

```python
# 5. Evaluate Datasets Based on String Distance Criteria
# Jaro-Winkler Similarity Distance: 0 = Exact Match, 1 = No Similarity

evaluation_config = RunEvalConfig(
    evaluators=[
        # You can define an arbitrary criterion as a key: value pair in t
        "string_distance",
        # Or to customize the distance metric:
        # RunEvalConfig.StringDistance(distance="levenshtein", normalize_
    ]
)

run_on_dataset(
    client=client,
    dataset_name="Elementary Animal Questions",
    llm_or_chain_factory=llm,
    evaluation=evaluation_config,
)
```

```
View the evaluation results for project 'bold-increase-22' at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17/compare?selectedSessions=73b730e7-e
1ba-4386-9bec-816bddf936fe

View all tests for Dataset Elementary Animal Questions at:
https://smith.langchain.com/o/3da9d385-1fd3-5914-b396-e10c2a27fd76/dataset
s/dd2191d5-f6a6-4677-b5b7-622fb131ec17
```

```
--------------------------------------------------------------------------
-
ModuleNotFoundError                       Traceback (most recent call las
t)
File /opt/anaconda3/lib/python3.11/site-packages/langchain/evaluation/stri
ng_distance/base.py:29, in _load_rapidfuzz()
     28 try:
---> 29     import rapidfuzz
     30 except ImportError:

ModuleNotFoundError: No module named 'rapidfuzz'

During handling of the above exception, another exception occurred:

ImportError                               Traceback (most recent call las
t)
Cell In[23], line 13
      1 # 5. Evaluate Datasets Based on String Distance Criteria
      2 # Jaro-Winkler Similarity Distance: 0 = Exact Match, 1 = No Simila
rity
      4 evaluation_config = RunEvalConfig(
      5     evaluators=[
      6         # You can define an arbitrary criterion as a key: value pa
ir in the criteria dict
   (...)
     10     ]
     11 )
---> 13 run_on_dataset(
     14     client=client,
     15     dataset_name="Elementary Animal Questions",
     16     llm_or_chain_factory=llm,
     17     evaluation=evaluation_config,
     18 )

File /opt/anaconda3/lib/python3.11/site-packages/langchain/smith/evaluatio
n/runner_utils.py:1368, in run_on_dataset(client, dataset_name, llm_or_cha
in_factory, evaluation, dataset_version, concurrency_level, project_name,
project_metadata, verbose, revision_id, **kwargs)
   1360     warn_deprecated(
   1361         "0.0.305",
   1362         message="The following arguments are deprecated and "
   (...)
   1365         removal="0.0.305",
   1366     )
   1367 client = client or Client()
-> 1368 container = _DatasetRunContainer.prepare(
   1369     client,
   1370     dataset_name,
   1371     llm_or_chain_factory,
   1372     project_name,
   1373     evaluation,
   1374     tags,
   1375     input_mapper,
   1376     concurrency_level,
   1377     project_metadata=project_metadata,
   1378     revision_id=revision_id,
   1379     dataset_version=dataset_version,
   1380 )
   1381 if concurrency_level == 0:
   1382     batch_results = [
   1383         _run_llm_or_chain(
   1384             example,
   (...)
   1389         for example, config in zip(container.examples, container.c
onfigs)
   1390     ]

File /opt/anaconda3/lib/python3.11/site-packages/langchain/smith/evaluatio
n/runner_utils.py:1188, in _DatasetRunContainer.prepare(cls, client, datas
et_name, llm_or_chain_factory, project_name, evaluation, tags, input_mappe
r, concurrency_level, project_metadata, revision_id, dataset_version)
   1186     run_metadata["revision_id"] = revision_id
   1187 wrapped_model = _wrap_in_chain_factory(llm_or_chain_factory)
-> 1188 run_evaluators = _setup_evaluation(
   1189     wrapped_model, examples, evaluation, dataset.data_type or Data
Type.kv
   1190 )
   1191 _validate_example_inputs(examples[0], wrapped_model, input_mapper)
   1192 progress_bar = progress.ProgressBarCallback(len(examples))

File /opt/anaconda3/lib/python3.11/site-packages/langchain/smith/evaluatio
n/runner_utils.py:433, in _setup_evaluation(llm_or_chain_factory, example
s, evaluation, data_type)
    431         run_inputs = chain.input_keys if isinstance(chain, Chain)
else None
    432         run_outputs = chain.output_keys if isinstance(chain, Chai
n) else None
--> 433     run_evaluators = _load_run_evaluators(
    434         evaluation,
    435         run_type,
    436         data_type,
    437         list(examples[0].outputs) if examples[0].outputs else Non
e,
    438         run_inputs,
    439         run_outputs,
    440     )
    441 else:
    442     # TODO: Create a default helpfulness evaluator
    443     run_evaluators = None

File /opt/anaconda3/lib/python3.11/site-packages/langchain/smith/evaluatio
n/runner_utils.py:618, in _load_run_evaluators(config, run_type, data_typ
e, example_outputs, run_inputs, run_outputs)
    614     input_key, prediction_key, reference_key = _get_keys(
    615         config, run_inputs, run_outputs, example_outputs
    616     )
    617 for eval_config in config.evaluators:
--> 618     run_evaluator = _construct_run_evaluator(
    619         eval_config,
    620         config.eval_llm,
    621         run_type,
    622         data_type,
    623         example_outputs,
    624         reference_key,
    625         input_key,
    626         prediction_key,
    627     )
    628     run_evaluators.append(run_evaluator)
    629 custom_evaluators = config.custom_evaluators or []
```

```
File /opt/anaconda3/lib/python3.11/site-packages/langchain/smith/evaluatio
n/runner_utils.py:530, in _construct_run_evaluator(eval_config, eval_llm,
run_type, data_type, example_outputs, reference_key, input_key, prediction
_key)
    528     if not isinstance(eval_config, EvaluatorType):
    529         eval_config = EvaluatorType(eval_config)
--> 530     evaluator_ = load_evaluator(eval_config, llm=eval_llm)
    531     eval_type_tag = eval_config.value
    532 elif isinstance(eval_config, smith_eval_config.EvalConfig):

File /opt/anaconda3/lib/python3.11/site-packages/langchain/evaluation/load
ing.py:147, in load_evaluator(evaluator, llm, **kwargs)
    145     return evaluator_cls.from_llm(llm=llm, **kwargs)
    146 else:
--> 147     return evaluator_cls(**kwargs)

File /opt/anaconda3/lib/python3.11/site-packages/langchain_core/load/seria
lizable.py:120, in Serializable.__init__(self, **kwargs)
    119 def __init__(self, **kwargs: Any) -> None:
--> 120     super().__init__(**kwargs)
    121     self._lc_kwargs = kwargs

File /opt/anaconda3/lib/python3.11/site-packages/pydantic/main.py:339, in
pydantic.main.BaseModel.__init__()

File /opt/anaconda3/lib/python3.11/site-packages/pydantic/main.py:1102, in
pydantic.main.validate_model()

File /opt/anaconda3/lib/python3.11/site-packages/langchain/evaluation/stri
ng_distance/base.py:77, in _RapidFuzzChainMixin.validate_dependencies(cls,
values)
     66 @root_validator
     67 def validate_dependencies(cls, values: Dict[str, Any]) -> Dict[st
r, Any]:
     68     """
     69     Validate that the rapidfuzz library is installed.
     70
   (...)
     75         Dict[str, Any]: The validated values.
     76     """
---> 77     _load_rapidfuzz()
     78     return values

File /opt/anaconda3/lib/python3.11/site-packages/langchain/evaluation/stri
ng_distance/base.py:31, in _load_rapidfuzz()
     29     import rapidfuzz
     30 except ImportError:
---> 31     raise ImportError(
     32         "Please install the rapidfuzz library to use the FuzzyMatc
hStringEvaluator."
     33         "Please install it with `pip install rapidfuzz`."
     34     )
     35 return rapidfuzz.distance

ImportError: Please install the rapidfuzz library to use the FuzzyMatchStr
ingEvaluator.Please install it with `pip install rapidfuzz`.
```

```
In [ ]:
```