BA-bench

Business Analytics Benchmark for GenAI Agents

Contribution Data Format

We aim to structure each sample (question-level) in our BA-bench using the following format:

```
{
    'unique_id': str,  # Unique identifier for each sample
    'question': str,  # A business-related question related to the data
    'data_file': str,  # The data file the question relates to
    'doc_file': str,  # The document file the question relates to
    'ground_truth': str,  # The answer to the question
    'data_domain': str,  # The domain the data belongs to (e.g., finance, education)
    'analysis_type': str,  # The type of question, optional: ["Structure problems", "Unst
    'origin_from': list[str],  # Source of the question, e.g., ['benchmark name', 'questio
    'additional_information': dict[str, str],  # Additional information such as code, sta
}
```

Analysis Type Definitions:

- **Structure problems**: The answer is structured (e.g., numerical, categorical).
  Example: A sample from *StatQA* where the `ground_truth` is a column and analysis method. If the agent's answer matches them, it is considered correct.
- **Unstructured problems**: The answer is unstructured (e.g., text-based).
  Example: A sample from *InsightBench* where the `ground_truth` is an insight. Semantic similarity must be considered.
- **Chart problems**: The answer is a plot.
  Example: A sample from *InsightBench* where the intermediate answer is a plot. The agent must write code to generate it.

Example Formats

StatQA Format

```
{
    'dataset': 'Dataset for Admission in the University',
    'refined_question': 'Is the variability in GRE scores not significantly different fro
    'relevant_column': '[{"column_header": "GRE Score", "is_strata": false, "is_control":
    'results': '[{"method": "Mood Variance Test", "result": "{\"stat\": 0.0, \"p value\":
    'ground_truth': '{"columns": ["GRE Score", "LOR"], "methods": ["Mood Variance Test", '
    'task': 'Variance Test',
    'difficulty': 'hard',
    'domain': 'Education & Student Performance'
}
```

InfAgent-DAbench Format

```
{
    'id': 0,
    'question': 'Calculate the mean fare paid by the passengers.',
    'concepts': ['Summary Statistics'],
    'constraints': "Calculate the mean fare using Python's built-in statistics module or a
    'format': '@mean_fare[mean_fare_value] where "mean_fare_value" is a floating-point nu
    'file_name': 'test_ave.csv',
    'level': 'easy',
    'domain': 'Tourism'
}
```

InsightBench Format

```
{
    "data_type": "descriptive",
    "insight": "The Hardware incidents are significantly higher in volume than others",
    "insight_value": {
        "x_val": "Hardware",
        "y_val": 335
    },
    "plot": {
        "plot_type": "histogram",
        "title": "Incidents by Category",
        "x_axis": {
            "name": "Category",
            "value": ["Hardware", "Software", "Network", "Inquiry / Help", "Database"],
            "description": "This represents the different categories of incidents."
        },
        "y_axis": {
            "name": "Number of Incidents",
            "value": [336, 41, 51, 32, 40],
            "description": "This represents the number of incidents in each category."
        },
        "description": "The histogram displays the distribution of incidents across differ
    },
    "question": "What is the distribution of incidents across all categories?",
    "actionable_insight": "Since the Hardware category has the highest number of incidents
    "code": "plot = df.groupby(\"category\").size().plot(kind=\"barh\", color=sns.palettes
}
```

Potentioal Benchmark

1. InsightBench

```
@article{sahu2024insightbench,
  title={Insightbench: Evaluating business analytics agents through multi-step insight gen
  author={Sahu, Gaurav and Puri, Abhay and Rodriguez, Juan and Abaskohi, Amirhossein and C
  journal={arXiv preprint arXiv:2407.06423},
  year={2024}
}
```

2. Infiagent-dabench

```
@article{hu2024infiagent,
  title={Infiagent-dabench: Evaluating agents on data analysis tasks},
  author={Hu, Xueyu and Zhao, Ziyu and Wei, Shuang and Chai, Ziwei and Ma, Qianli and Wang
  journal={arXiv preprint arXiv:2401.05507},
  year={2024}
}
```

3. StatQA

```
@article{zhu2024large,
  title={Are Large Language Models Good Statisticians?},
  author={Zhu, Yizhang and Du, Shiyin and Li, Boyan and Luo, Yuyu and Tang, Nan},
  journal={arXiv preprint arXiv:2406.07815},
  year={2024}
}
```

4. DSBench

```
@article{jing2024dsbench,
  title={DSBench: How Far Are Data Science Agents to Becoming Data Science Experts?},
  author={Jing, Liqiang and Huang, Zhehui and Wang, Xiaoyang and Yao, Wenlin and Yu, Wenha
  journal={arXiv preprint arXiv:2409.07703},
  year={2024}
}
```

Collaboration Work

- one person to process InsightBench
- one person to process StatQA and Infiagent-dabench