ESG Categorization Project Guidelines

We've been working on designing a prompt engineering framework to categorize fund strategies based on their alignment with six ESG categories: Apply Exclusions, Limit ESG Risk, Seek ESG Opportunities, Practice Active Ownership, Target Sustainability Themes, and Assess Impact. The data is in Wanqi's folder labeled $results - together.csv$.

Goal

The ultimate objective is to assign a score between 0 and 1 for each of the six categories, based solely on the text found in the 'strategy' field of a fund. Scores reflect how much the fund strategy aligns with each ESG category:

- 1 indicates a clear and dominant focus,
- 0 means the category is not relevant or mentioned,
- Intermediate values represent proportional alignment.

Key Developments

Prompt Refinement: We've iteratively refined the prompt through multiple versions to clarify the ESG category definitions and improve how GPT-4 interprets fund strategies. The final prompt explicitly:

- Provides full, unambiguous definitions for each ESG category.
- Instructs the model to assign 0 if a category is not explicitly mentioned.
- Prevents score inflation when only vague ESG references are present.
- Clarifies when to assign intermediate scores and enforces score logic (e.g., only one category should score 1 unless others are clearly secondary).

Prompt Validation: We've tested this prompt on several example fund strategies, including:

- Clear multi-category strategies.
- Long strategies with only small ESG references.

- Ambiguous or indirect ESG language. The current version is now able to assign logical, grounded scores and handles category distinctions well.

Common Issues We Solved:

- Early versions often assigned scores ¿0 to all categories just because the text "sounded ESG-like."
- The model sometimes inflated scores for categories like "Assess Impact" even when no measurement was mentioned.
- We clarified how the model should handle diluted ESG references in long strategy texts.

Next Steps

- Use the final prompt (stored in our shared drive/document) for GPT API-based batch testing on our dataset of ∼19,815 fund strategies.
- Evaluate outputs for consistency, especially edge cases where ESG language is subtle or buried.
- Document examples where the scoring seems off and adjust prompt wording if needed.
- Optionally explore semi-automated labeling using a hybrid of GPT output + logistic regression or Naive Bayes as a validation layer (this was an earlier idea we paused).

Key Tip

Remember: no category should get a score unless it's clearly represented in the strategy text. GPT may still hallucinate if the prompt is vague.

Final Prompt

Please evaluate the following fund strategy across the six ESG categories described below. Assign a score between 0 and 1 for each category based on its relevance to the strategy:

- A score of 1 indicates that the category is the clear and dominant focus of the strategy.
- Scores between 0.1 and 0.9 should be assigned proportionally, reflecting partial but meaningful alignment.
- A score of 0 must be assigned if the category is not explicitly mentioned or if there is no clear evidence of alignment.

Important Notes:

- Do not assign a score above 0 unless the strategy text clearly refers to that category.

- A fund may reference ESG in general, but this does not mean all categories are relevant.
- In long strategy texts, ESG-related statements may be few. Only score based on the specific content, not the overall length or tone.
- It is acceptable for all scores to be 0 if the strategy is not ESG-focused.

ESG Categories and Scoring Guidelines

Apply Exclusions

Excluding sectors, companies, or practices that are harmful or misaligned with sustainability values (e.g., tobacco, fossil fuels, gambling, human rights violators).

- Score 1.0 if exclusions are the primary focus.
- Score proportionally for narrower or partial exclusions.
- Score 0.0 if no exclusions are mentioned.

Limit ESG Risk

Managing ESG risks using ESG data/ratings (e.g., avoiding companies with high climate risk or social governance issues).

- Score 1.0 if ESG risk limitation is the main focus using data/analysis.
- Score proportionally if risk is discussed but not using ESG-specific tools.
- Score 0.0 if no mention of ESG risk management.

Seek ESG Opportunities

Actively investing in sustainability leaders or companies improving their ESG performance to gain competitive advantage.

- Score 1.0 if ESG opportunity-seeking is dominant.
- Score proportionally if sustainability is a factor but not the main strategy.
- Score 0.0 if there's no mention of proactively seeking ESG-positive firms.

Practice Active Ownership

Shareholder engagement or stewardship (e.g., proxy voting, dialogues with companies to improve ESG practices).

- Score 1.0 if the strategy emphasizes these actions.
- Score proportionally for minor references.
- Score 0.0 if there's no mention of active ownership.

Target Sustainability Themes

Explicit focus on themes like climate action, clean energy, social equity, biodiversity, SDGs, etc.

- Score 1.0 if the strategy centers around one or more themes.
- Score proportionally if themes are present but not dominant.
- Score 0.0 if no themes are clearly mentioned.

Assess Impact

Measuring, reporting, or targeting quantifiable social or environmental impact.

- Score 1.0 if impact measurement is a core feature.
- Score proportionally if it's implied but not detailed.
- Score 0.0 if there is no mention of tracking or reporting impact.

Test Against:

- One-category examples.
- Multi-category examples.
- Long strategies with diluted ESG references.
- Ambiguous or vague ESG language.

Next Steps

Tweak the prompt and test it more. Use GPT 4.5 for testing as well as o3-mini.
Can test against other AI assistants.
Make use of the current free month.

Automation

Clean up the data a bit. Truncate to the first 1000-1500 tokens. Re-work prompt based on the data.

Sample prompt based on the data:

> You will now evaluate the following fund strategy based on the extent to which it aligns with each of the six ESG categories. The strategy may be long and contain non-ESG content — focus only on identifying ESG-related goals, approaches, or implications, even if ESG terms are not explicitly mentioned.
>
> Assign a score from 0.0 to 1.0 for each ESG category. A score of 1.0 should only be used if a category is the dominant theme. If a category is not present, assign a score of 0. Do not assign scores based on financial, operational, or legal language unless there is a clear ESG-related motivation behind it.

Categories:

1. Apply Exclusions: Avoiding specific sectors, companies, or practices considered harmful or unethical (e.g., arms, fossil fuels, human rights violators), even if not labeled as “exclusions.”

2. Limit ESG Risk: Strategies to manage material ESG risks — e.g., avoiding firms with poor environmental or social performance, integrating ESG risk into decision-making.

3. Seek ESG Opportunities: Actively investing in companies or sectors that are ESG leaders or are improving ESG practices to gain competitive advantage.

4. Practice Active Ownership: Engaging with portfolio companies to improve ESG practices — includes dialogue, proxy voting, or ESG-focused shareholder resolutions.

5. Target Sustainability Themes: Thematic focus on issues such as climate action, diversity, health, education, clean energy, etc. These may reflect alignment with U.N. SDGs or other frameworks.

6. Assess Impact: Any attempt to measure, report, or optimize the environmental or social impact of the fund’s investments.

Instructions:

- Score based on the intended meaning, not the presence of keywords.
- If ESG is only partially present, assign proportionally lower scores.
- If no category is clearly present, assign all scores as 0.

Now score the following strategy:

Python Script for ESG Categorization Using GPT API

```
import openai
import pandas as pd
import time

# Load your CSV
df = pd.read_csv("results_together.csv")

# Change ’investment’ if needed
df[’strategy_text’] = df[’investment’]

# Define final refined ESG evaluation prompt
prompt_prefix = """
You are an ESG fund strategy evaluator. You will receive a fund strategy and evaluate it acros
Focus only on ESG-related content. Do not assign positive scores based on financial or legal c
```

```
Categories:

1. Apply Exclusions: Excluding sectors, companies, or practices that are harmful or misaligned
   If the strategy avoids companies for ethical or ESG reasons, this counts|even if the term "e

2. Limit ESG Risk: Incorporating ESG risk factors into decision-making or avoiding firms with p
   This includes any use of ESG data to manage downside risks.

3. Seek ESG Opportunities: Investing in companies that are ESG leaders or improving ESG practic
   Includes positive screening or best-in-class investing.

4. Practice Active Ownership: Engaging with companies to improve ESG performance via proxy voti
   Must be clear about engagement intent.

5. Target Sustainability Themes: Thematic focus on areas like climate action, health, diversity
   May reference alignment with U.N. SDGs.

6. Assess Impact: Explicitly measuring or reporting the environmental/social impact of investme
   If not explicitly measured, this category should be scored 0.

Scoring Instructions:
- Assign a score from 0.0 to 1.0 for each category.
- A score of 1.0 means the category is a dominant focus.
- A score of 0.0 means it is not mentioned or implied.
- Use intermediate scores (e.g., 0.2, 0.5, 0.7) if a category is partially represented.
- If no ESG category applies at all, assign all scores as 0.
- Be accurate and concise in your explanations.

Now, evaluate the following fund strategy:
"""

# Replace with OpenAI API key
openai.api_key = "api-key-here"

results = []

for idx, row in df.iterrows():
    strategy = row['strategy_text']
    prompt = prompt_prefix + "\n" + strategy.strip()

    try:
        response = openai.ChatCompletion.create(
            model="gpt-4",
            messages=[{"role": "user", "content": prompt}],
            temperature=0.2
        )
        output = response.choices[0].message["content"]
        results.append(output)
        print(f"Processed row {idx+1}/{len(df)}")
        time.sleep(1.5)  # Respect rate limits
    except Exception as e:
        print(f"Error at row {idx+1}: {e}")
        results.append(f"Error: {e}")

# Save outputs
df['esg_scores'] = results
df.to_csv("esg_scores_output.csv", index=False)
print("Saved ESG scores to esg_scores_output.csv")
```