Annotation Guideline for Evaluating the Cognitive Levels of Questions in Financial Earnings Calls

1. Introduction

1.1. Purpose

The purpose of our annotation guideline is to standardize the process of annotating coherent questions in Q&A sessions of the earnings call based on their cognitive demand according to Bloom's Taxonomy, a pedagogical framework. Benjamin Bloom, an American psychologist, discovered that the majority of the test questions only require students to think at the lowest level - recall of factual information. Thus Bloom's Taxonomy was originally developed as a guideline to classify levels of intellectual complexity required in learning to improve educational quality. We will first introduce a modified version of Bloom's Taxonomy, creating four intellectual categories: knowledge, comprehension, application/methodology, and analysis/evaluation. The task is defined as using the tags to assign a cognitive level from {1, 2, 3, 4} on a question based on its cognitive complexity and the level of thinking it tries to induce from its respondent.

1.2. Scope

We try to expand beyond the context of education and apply Bloom's ideas to the financial industry, specifically on questions collected from the transcripts of financial earning calls. This task should provide evidence on the **complexity and depth of inquiries** from financial analysts, aiding institutions and investors in developing a more thorough understanding of the content covered in the Q&A section of the call. This guideline is developed without direct references to the financial sector, thereby ensuring its applicability to a broad range of questions. Potential areas of application include Q&A sections from conferences, workshops, and webinars.

1.3 Approach

The annotator should determine the cognitive levels based on the tags outlined in Section 2.1. Note that a single question might include parts or even sub questions that fall under distinct tags.These tag(s) are then combined with defined bullet points as evaluation metrics. Questions are assigned into cognitive levels 1 through 4 from least to most cognitively demanding. Annotators should holistically evaluate the related

cognitive tags when determining a final score. This complexity suggests the need for a cognitive scoring system that extends to the idea of cognitive tagging.

2. Annotation Guidelines for Cognitive Domains

2.1 Tag formulation

Tag(1) Knowledge:

- Questions that require the identification or recall of factual information. These questions typically ask for specific details, definitions, or descriptions.
- Verb examples: arrange, define, duplicate, label, list, memorize, name, order, recognize, relate, recall, repeat, reproduce, state

Tag(2) Comprehension:

- Questions that are about the understanding and interpretation of some specific organization, facts, and ideas. The target of interest is mostly on an existing matter.
- Verb examples: classify, describe, discuss, explain, express, identify, indicate, locate, recognize, report, restate, review, select, translate

Tag(3) Application and Methodology:

- Questions about the use and application of methods, rules, and principles. This involves selecting and applying these methods to solve problems or achieve objectives via applying acquired knowledge in new ways and different settings.
- Verb examples: design, apply, implement, select, utilize, employ, perform, plan, operate, schedule, develop, solve, use, adapt

Tag(4) Analysis and Evaluation:

- Questions about the evaluation of information and creating relevant value judgments. These questions can sometimes involve a segmentation process where information is broken up into components and analyzed.
- Verb examples: analyze, appraise, categorize, compare, contrast, criticize, differentiate, discriminate, distinguish, examine, experiment, appraise, argue, assess, attach, choose compare, defend, estimate, judge, predict, rate, core, select, support, value, evaluate

Note: All verb examples are cited from here.

2.2 Level Classification

Level 1 (Fact/Statistics):

Questions only consist of segments that can be categorized as Tag (1) with Tag (1) being the dominant segment. An inquiry on factual recall

- Usually has a limited amount of **straightforward answers** (yes/no, correct/incorrect, true/false)
    - e.g. “Is there a number for coverage rate?”
- A **truth value can be assigned** to the induced answer based on facts
    - e.g. “So is it a safe assumption to say that you've more than hit your target of 40 new products that generate over $1 million in sales?”

Level 2 (Clarification/Explanation/No Inference):
Questions consist of segments that can be categorized as Tag (2) with Tag (2) being the dominant segment

- Potential answers to these questions are in a **certain direction** – they can be worded differently but lead to the same idea
    - e.g. “You brought down the guidance for comps in Q2 and Q3, could you talk about your thoughts behind doing that?”
- Potential answers provide follow-up **details via the respondent’s clarification**
    - e.g. “Is that just more administrative to put that out there, or is there something I should read into it?”
- Can motivate a set of **explanations** or rationales behind a fact or situation
    - e.g. “And just in terms of this refinancing, given the size of your cash balance, why not just pay down the debt outright?”
- Note: there should be **no inference** to an extended matter in the questions. The question should not be constructed via the use of existing information or logical reasoning

Level 3 (Methodology/Inference):
1）Questions consist of segments that can be categorized as Tag (3)

- Contain requests for the **methodology** for issues (sometimes hypothetical)
    - e.g. “How would you lock in the liquid asphalt cost for the rest of the year?”
- **Predictions** of specific objects (quantity, trend) based on the acquired knowledge
    - e.g. “Do you expect operating profit in sweeteners to hold steady in 2014 given the fructose pricing?”

2）Questions consist of segments that can be categorized as Tag (4), potentially composed of requests for specific evaluations based on the set criteria or hypothesis.

- Questions generally ask for more **specific** and **focused evaluations** that can be tackled by using evidence and reasoning
    - e.g. Can you just review for us where silicone hydrogel daily as a category is today just from a market standpoint? And where do you think it can go over time?
- Simple **comparison** and **analysis** that may depend on given criteria
    - e.g. “And I’m just wondering, the variability of cost and labor and some of the other things, what does margin look like in the 3% to 5% comp versus a 7% to 9% comp?”

    - e.g. “What’s the advantage of the product you introduced, especially considering potential technological advancements and shifts in consumer preferences?”
- Can refer to a **hypothetical or external scenario** to provoke simple analysis
    - e.g. “If that market took off, would you consider developing your own spark-ignited 13-liter?”

Level 4: (Open-ended Question/Strategy/Plan/Contradiction)
Questions consist of segments that can be categorized as Tag (4), potentially composed of open-ended inquiries with different variations and alternative answers.

- Contains **synthesized and composite information** that requires a higher level of understanding
    - e.g. “You said it was a mix between 3D tomo and Panther pull-through. Can you talk about the mix between both of those back, forward, and how is that going to impact 2014 and how we think about guidance going there?”
- Questions generally ask for **comprehensive and open-ended evaluations** that are supported by evidence and reasoning
    - e.g. “On hep C, can you outline the development plan for the Idenix nuc and when we should next see data on this asset? Then perhaps, while we are on the nuc, can you remind us of what gave you comfort to acquire this, in light of the turbulent history with that class?”
- Addressing **Contradictions** with Evidence-Based Reasoning
    - e.g. “How does the math work when you are basically increasing your tonnes by 60% but the production is only going up by 15% or 20% what am I missing?”
    - e.g. “What is causing it to be lower sequentially if trends are getting much better?”

2.3 Supplementary Note

Given that a question can consist of segments that are associated with various tags, we define **dominance** and **existence** for a more robust guideline.

- For tag (1) and tag (2), we determine the **dominant tag** by evaluating the prevalence [proportional to the length of the question] for tag (1) and tag (2). If both tags take up equivalent portions of the question, tag (2) becomes the dominant tag.
    - For example:
      “I was wondering if you are seeing any of the services yet in your markets, and what are your thoughts as far as the services being a competitive threat to wireline?”
    - The segment before the comma can be categorized as tag (1) where the one after is one-word longer and can be categorized as tag (2). Thus the question is classified as level 2 due to tag (2) being dominant.

- For tag (3) and tag (4), more emphasis is put on **existence**. The highest-ranking tag is identified and used as a baseline to determine the levels of thinking.
    - For example:

"You said it was a mix between bond A and bond B, can you talk about the mix between both of those back, forward, and how that is going to impact 2014 and how we think about guidance going there?"

- The first segment, "you said it was a mix…" would be classified as tag (1), "can you talk about the mix…" would be classified as tag (2), "how that is going to impact 2014" would be classified as tag (3), and "how we think about guidance going there" would be classified as tag (4). Given that tag(4) is the highest-rank tag, we default to a classification between Level 3 and Level 4. Coupled with the composite bullet point, this question is finally classified as level 4.