Capstone 2026

High-Impact Business Constructs for Vertical AI

Seven Project Proposals

Methodological Foundation: Ringel (2023), "Creating Synthetic Experts with Generative Artificial Intelligence"

Anchor Validation: Ludwig et al. (2026), "Extracting Consumer Insight from Text" (arXiv:2602.15312)

Methodological Playbook

All seven projects follow the *synthetic experts* pipeline introduced by Ringel (2023). The core idea is to use generative AI (frontier LLMs such as GPT-5.4, Grok-4, Deepseek, Gemini 3.1 or Claude Sonnet 4.6) to produce training labels for theoretically grounded constructs where expert annotations are scarce or expensive, then fine-tune an open-weights model (e.g., RoBERTa Large or a domain-adapted BERT variant) to approximate the GenAI labeler at low marginal cost and high throughput. Ringel (2023) demonstrates this approach for marketing-mix variable detection in microblogs, shows that GenAI labels can align well with expert labels, and introduces *synthetic twins* as a privacy-preserving text-replication technique for sharing training data without exposing original user-generated content.

A second anchor reference is Ludwig et al. (2026; arXiv:2602.15312), who build LX, a fine-tuned model trained on consumer-authored text labeled with 16 self-reported emotions and four evaluation constructs (trust, commitment, recommendation intent, and sentiment). LX outperforms strong baselines on both survey open-ends and online reviews, exemplifying the end-state each project targets: a reusable, validated, open classifier that other researchers and practitioners can deploy.

Exemplary Pipeline Steps (Common to All Projects)

- **Step 1 — Construct definition and rubric.** Define the target construct with clear theoretical grounding, operationalize it as a classification task, and write a detailed labeling rubric with definitions, anchor examples, and boundary cases. The rubric is the single most important artifact for label quality. Decide on the unit of analysis (sentence, paragraph, document, etc.)
- **Step 2 — Construct Train and Holdout Samples.** Collect your data and make sure you have sufficient examples to work with (I recommend 50k+). Ensure class/label balance and sufficient representation in holdout and in train sample. Use oversampling where necessary. At a minimum, have 1.2K holdout and 20k training examples.
- **Step 3 — Human labeling and adjudication.** Label the holdout sample as a team (consider oversampling disagreement cases and rare classes). Consider using disagreement-mining and LLM-as-judge adjudication for hard cases (Ma et al., 2026). Compute inter-rater reliability / Krippendorff's alpha and correct systematic biases.
- **Step 4 — GenAI labeling and Model Benchmarking.** Label the holdout sample with genAI models (use chat and reasoning models). Revisit your system prompt where necessary to get sufficient performance of at least one model that you will use to label the training data. If your F1 is below .9, then your training data will have too much label noise.
- **Step 5 — GenAI labeling at scale.** Use the frontier LLM that performed best on the holdout (with the rubric in the system prompt) to label a large training corpus (typically at least 20K text units). Consider using multi-model consensus (cf. EvasionBench; Ma et al., 2026) to improve reliability.

- **Step 6 — Fine-tune an open-weights model.** Train a smaller, efficient model (RoBERTa, DestillBERT, specialized BERT, etc.) to approximate the GenAI model. This distillation step is what makes the measure deployable at scale and at near-zero marginal cost.
- **Step 7 — Validation.** Evaluate on (a) the holdout human-coded set of at least(!) 1.2k examples, (b) robustness splits by time period, industry, and text domain, (c) calibration metrics for probabilistic outputs, and (d) downstream/nomological validity tests (i.e., does the predicted construct relate to expected real-world outcomes).
- **Step 8 — Document.** Document everything and create charts that show model comparison in consistency, performance, cost and timing. Ideally, include an external benchmark approach (i.e., previous best practice).

Take a close look at this exemplary pipeline that I vibe coded for you:
http://ringel.ai/UNC/2026/helpers/Ringel_2026_VerticalAI_Capstone_Pipeline_Example.zip

This will give you a better impression of the key and minimal steps you need to take.

Labeling Enhancements from Recent Literature

Two recent papers offer concrete enhancements that every team can adopt. First, EvasionBench (Ma et al., 2026) introduces a scalable multi-model consensus strategy: generate labels from multiple LLMs independently, mine disagreements, and use an LLM-as-judge to adjudicate hard cases. This is directly compatible with the Ringel (2023) pipeline and is particularly useful for abstract or subjective constructs. Second, the ESG-Activities paper (Kiefer et al., 2025; arXiv:2502.21112) explicitly demonstrates gains from fine-tuning on a mixture of human-curated and synthetic data, confirming that a blended data strategy improves classification performance in ESG text. However, be careful with synthetic data! If done poorly, you can inadvertently create biases and overfit to these data such that your model generalizes poorly!

Recommended Evaluation Package

Each project should report: (a) macro-F1 and per-class F1 on a held-out human-labeled test set; (b) Krippendorff's alpha between the GenAI labeler and human annotators; (c) calibration plots (reliability diagrams) for probabilistic outputs; (d) robustness checks across time periods, industries, and text sources; and (e) ideally one downstream validity test demonstrating that the classifier's predictions correlate with theoretically expected outcomes (e.g., market reactions, operational performance, regulatory actions).

Exemplary Presentation

Take a look at the Class 21 materials for a blueprint:

Smarter, cheaper, greener: Vertical AI for business analytics

http://www.ringel.ai/UNC/2026/BUSI488/Class21/Ringel_488-2026_Class21.pdf

Project 1: Consumer Emotions and Evaluations (LX-Style Extension)

Why It Matters

Understanding what consumers feel and how they evaluate products is foundational to marketing strategy, brand management, and customer experience design. Emotions drive purchase decisions, word-of-mouth, and loyalty; evaluation constructs such as trust, commitment, and recommendation intent are direct levers for customer lifetime value. Yet most text-analytic emotion tools are trained on generic corpora and validated against lexicons rather than self-reported ground truth, leaving a significant measurement gap. Ludwig et al. (2026) address this gap directly by building LX, a fine-tuned language model trained on consumer text paired with self-reported labels for 16 discrete emotions (e.g., joy, anger, surprise, guilt, pride, gratitude) and four evaluation dimensions (trust, commitment, recommendation intent, and overall sentiment). LX outperforms GPT-4, dictionary-based methods, and prior supervised models on both survey open-ends and online product reviews, establishing a new standard for consumer-text measurement.

Theoretical Core

The project is grounded in appraisal theories of emotion (Roseman, 1991; Smith & Ellsworth, 1985), which hold that distinct emotions arise from cognitive evaluations of events along dimensions such as pleasantness, agency, certainty, and control. In marketing, the consumption-emotion literature (Richins, 1997) establishes that consumers experience a structured set of emotions during and after product use, and that these emotions predict satisfaction, complaint behavior, and repurchase. Ludwig et al. (2026) select their 16-emotion taxonomy by intersecting appraisal-theory predictions with marketing-relevant consumption experiences, ensuring that each emotion class has both theoretical justification and practical salience.

The four evaluation constructs—trust, commitment, recommendation intent, and sentiment—are drawn from relationship marketing theory (Morgan & Hunt, 1994) and the Net Promoter framework (Reichheld, 2003). Trust and commitment are established mediators of customer retention; recommendation intent captures advocacy behavior; sentiment provides a general valence summary. Together, these constructs form a comprehensive "consumer insight" profile extractable from a single text.

Prior Approaches and Gaps

Prior approaches to emotion detection in consumer text fall into three categories: (1) lexicon-based methods (e.g., NRC Emotion Lexicon), which are transparent but miss context-dependent meaning, negation, and sarcasm; (2) supervised classifiers trained on annotator-labeled corpora (e.g., GoEmotions), which suffer from the gap between annotator perception and the author’s actual emotional state; and (3) zero-shot LLM prompting, which is flexible but expensive at scale and inconsistent across prompts. Ludwig et al. (2026) demonstrate that LX—a model fine-tuned on text where the same person who wrote the text also self-reported their

emotions—outperforms all three approaches, closing the "ground-truth gap" between what annotators infer and what consumers actually feel.

Construct Definition and Label Schema

The label schema comprises two tasks applied to each text unit (review paragraph or survey open-end):

Task A: Emotion classification. Multi-label assignment across 16 discrete emotions: joy, love, surprise, pride, gratitude, contentment, relief, hope, interest/excitement, anger, frustration, disappointment, sadness, fear/worry, guilt, and disgust. Each emotion is coded as present or absent, allowing for co-occurring emotions (e.g., a review expressing both gratitude and disappointment).

Task B: Evaluation constructs. Four separate ordinal scales: trust (low/medium/high), commitment (low/medium/high), recommendation intent (would not recommend / neutral / would recommend), and overall sentiment (negative/neutral/positive). These are scored per text unit independently of the emotion labels.

Label Schema Summary

Dimension	Classes / Labels	Granularity
Emotions (16)	Joy, Love, Surprise, Pride, Gratitude, Contentment, Relief, Hope, Interest, Anger, Frustration, Disappointment, Sadness, Fear, Guilt, Disgust	Multi-label per text unit
Trust	Low / Medium / High	Ordinal per text unit
Commitment	Low / Medium / High	Ordinal per text unit
Recommendation	Would not / Neutral / Would recommend	Ordinal per text unit
Sentiment	Negative / Neutral / Positive	Ordinal per text unit

Data Sources

The primary data sources are publicly available review corpora and, where possible, partner survey data with self-reported emotion labels. The Amazon Reviews 2023 corpus (McAuley Lab, Hugging Face) provides hundreds of millions of product reviews with star ratings across dozens of product categories. The Yelp Academic Dataset provides multi-million-review service/restaurant data with star ratings. For self-reported ground truth, the Ludwig et al. (2026) design—pairing open-ended text with a contemporaneous emotion/evaluation survey—could be replicated via a new survey panel or through a research partnership. ***However, this is likely beyond the capstone project!*** Hence, you need to create the ground truth as expert labelers by inferring them from the text. While this approach is not the same as Ludwig et al. (2026) who use the same consumers who wrote the text to report their emotions, it is okay to use with the capstone!

https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

https://s3-media2.fl.yelpcdn.com/assets/srv0/engineering_pages/e926cc12796d/assets/vendor/yelp-dataset-license.pdf
https://arxiv.org/abs/2602.15312

Feasibility Risks and Mitigation

The main risk is obtaining self-reported emotion labels at scale since reviews alone do not come with ground-truth emotion data. Mitigation: (1) use the Ringel (2023) synthetic-expert pipeline to generate high-quality GenAI labels for the large review corpora, validated against a smaller set of either self-reported labels collected via a survey panel OR use the team as human experts to infer the labels from the texts; (2) use Ludwig et al.'s published results as a benchmark to calibrate the GenAI labeler's alignment with self-report; and (3) apply synthetic-twins for any data-sharing. A secondary risk is class imbalance (some emotions like guilt or disgust are rare in product reviews); mitigate with stratified sampling and class-weighted loss during fine-tuning.

Expected Difficulty and Novelty

Difficulty: medium (replication and extension of a well-specified pipeline). Novelty: moderate to high, depending on the extension—e.g., cross-domain transfer (service vs. product), multilingual deployment, or integration with customer journey analytics.

Project 2: Earnings-Call Evasiveness Classifier (EvasionBench-Style)

Why It Matters

Whether executives answer analyst questions directly during earnings calls is a signal of transparency, information asymmetry, and managerial confidence. Evasive responses can predict subsequent earnings surprises, increased analyst uncertainty, and negative market reactions. Yet measuring evasiveness at scale has been stymied by the cost of expert annotation and the subtlety of the construct: evasion is not merely vagueness but a strategic communicative act in which the speaker substitutes an unasked question for the one posed. Ma, Lin, and Yang (2026) address this gap with EvasionBench, a benchmark dataset of earnings-call Q&A turns labeled at three evasion levels, supported by a multi-model consensus labeling framework and an open dataset on Hugging Face.

Theoretical Core

The construct is rooted in impression management theory (Leary & Kowalski, 1990) and strategic ambiguity in corporate communications (Eisenberg, 1984). Executives use evasion as a tool to manage information asymmetry while maintaining plausible deniability; the degree of evasion reflects a trade-off between legal caution, competitive sensitivity, and the reputational cost of perceived non-transparency. In accounting research, Hollander, Pronk, and Roelofsen (2010) show that managers' willingness to answer questions directly is informative about future performance, and that markets partially price evasion. EvasionBench (Ma et al., 2026) operationalizes evasion as a three-level construct—direct, partially evasive, and fully evasive—where "partially evasive" captures answers that address the topic but deflect from the specific question asked, and "fully evasive" captures outright topic-shifting or non-answers.

Prior Approaches and Gaps

Older approaches proxy evasiveness with linguistic features such as hedge-word counts, answer length, or generic uncertainty dictionaries (e.g., Loughran & McDonald, 2011). These proxies conflate evasion with legitimate hedging, technical complexity, and legal boilerplate. EvasionBench introduces a direct, construct-aligned measurement: human-annotated turn-level evasion labels, a multi-model labeling pipeline that mines disagreement between LLMs and uses an LLM-as-judge to adjudicate, and a publicly released dataset and model artifacts. The gap this project fills is deploying the EvasionBench framework within a Ringel (2023) fine-tuning pipeline to produce a lightweight, scalable classifier that can be applied to any earnings-call corpus.

Construct Definition and Label Schema

The unit of analysis is a Q&A turn pair: one analyst question and one executive response. The primary label is a single three-class ordinal variable:

Label	Definition	Anchor Example (Abbreviated)
Direct	The response substantively answers the specific question asked, with concrete information.	Q: What drove the margin decline? A: Three factors—raw material costs up 12%, freight surcharges, and a one-time write-down of...
Partially Evasive	The response addresses the general topic but deflects from the specific question or substitutes a different framing.	Q: What drove the margin decline? A: We're very focused on long-term margin improvement and expect to see benefits from our restructuring...
Fully Evasive	The response avoids the topic entirely, pivots to an unrelated subject, or provides a non-answer.	Q: What drove the margin decline? A: I'd like to highlight the strong revenue growth we saw in our cloud segment this quarter...

Optional multi-label extensions include evasion tactic type (pivot, denial, information overload, legal safe harbor, forward-referral to written materials) and whether the analyst pressed for a follow-up.

Data Sources

The EvasionBench dataset is publicly available on Hugging Face and provides pre-processed Q&A turn pairs with human annotations. For extending to a larger corpus, earnings-call transcripts are typically obtained from commercial providers such as S&P Capital IQ, Refinitiv (now LSEG), or Seeking Alpha. Institutional access via a business school subscription is the most common route. Note: transcript access must be confirmed before committing to a large-scale labeling effort.

https://huggingface.co/datasets/FutureMa/EvasionBench
https://github.com/IIIIQIIII/EvasionBench
https://arxiv.org/abs/2601.09142

Feasibility Risks and Mitigation

The primary risk is transcript licensing: commercial transcript providers restrict redistribution, and some restrict NLP analysis in their terms of service. Mitigation: (1) use EvasionBench's open dataset as the primary training seed; (2) release only derived labels and model weights, never raw transcript text; (3) confirm institutional access before scaling. A secondary risk is the confusion between "legal caution" and "evasion"—some answers are non-specific because of legitimate regulatory constraints. Mitigation: include a "legal safe harbor" tactic label and train the model to distinguish formulaic legal language from substantive evasion; human audit should specifically adjudicate this boundary.

Expected Difficulty and Novelty

Difficulty: medium (benchmark exists; pipeline is well-specified). Novelty: moderate (extension to Ringel-style distillation, tactic subtypes, and validation against market outcomes.

Project 3: Cybersecurity Governance and Incident Disclosure Quality (SEC-Aligned)

Why It Matters

Cybersecurity risk is now among the most financially material operational risks facing firms. In July 2023, the SEC adopted final rules (Release 33-11216) requiring registrants to disclose material cybersecurity incidents on Form 8-K (Item 1.05) and to describe their cybersecurity risk management, strategy, and governance annually on Form 10-K (new Item 1C). This regulatory mandate creates both a compliance obligation and a rich text corpus for analysis. Investors, boards, and regulators need tools to assess whether disclosures are substantive or boilerplate, whether governance structures are robust or ceremonial, and whether incident reports are timely and informative. No validated, construct-aligned classifier exists for this purpose.

Theoretical Core

The theoretical foundation draws on disclosure theory (Verrecchia, 2001) and regulatory compliance as an information-provision mechanism. The SEC's final rule specifies four content domains: (1) governance—board oversight and management roles in cybersecurity risk; (2) risk management processes—how the firm identifies, assesses, and manages cyber threats; (3) strategy integration—how cybersecurity is embedded in overall business strategy; and (4) incident disclosure—scope, timing, materiality, and remediation of specific incidents. The rule's structured requirements provide a natural taxonomy that maps directly to a multi-class classification task. Complementary literature on information security governance (e.g., Von Solms & Von Solms, 2004) and IT risk disclosure (e.g., Gordon, Loeb, & Sohail, 2010) provides additional theoretical scaffolding for the specificity dimension: disclosures vary from generic boilerplate to concrete, decision-useful descriptions with named technologies, quantified impacts, and specific timelines.

Prior Approaches and Gaps

Existing approaches largely rely on keyword dictionaries (e.g., counting mentions of "cybersecurity," "data breach," or "encryption") or topic models that identify cyber-related passages without classifying their content or quality. The gap is a structured classifier that maps disclosure paragraphs into the SEC rule's content categories and simultaneously measures specificity (how informative is the paragraph?) on an ordinal scale. The SEC rule itself essentially provides the labeling rubric, making this a high-feasibility project for GenAI labeling.

Construct Definition and Label Schema

The unit of analysis is the paragraph within Item 1C (10-K) or Item 1.05 (8-K). Two label dimensions are applied simultaneously:

Dimension 1: Content category. Single-label classification into one of five categories: (a) Board/governance oversight, (b) Management roles and responsibilities, (c) Risk management processes, (d) Strategy integration, (e) Incident disclosure (scope, timing, impact, remediation), and (f) None/other (paragraph does not contain cybersecurity-specific disclosure content).

Dimension 2: Disclosure specificity. Ordinal rating on a 4-point scale: (1) Generic boilerplate—could apply to any firm with no modification; (2) Sector-adapted—references industry but no firm-specific details; (3) Firm-specific—names technologies, teams, timelines, or processes unique to the firm; (4) Quantified and verifiable—includes specific metrics, dollar amounts, incident timelines, or third-party audit references.

Label Schema Summary

Dimension	Classes / Labels	Granularity
Content Category	Board oversight / Management roles / Risk mgmt processes / Strategy integration / Incident disclosure / None	Single-label per paragraph
Specificity	Generic boilerplate / Sector-adapted / Firm-specific / Quantified-verifiable	Ordinal per paragraph

Data Sources

All data is publicly available through SEC EDGAR. Form 10-K Item 1C filings are accessible via EDGAR full-text search and the EDGAR API. Form 8-K Item 1.05 incident disclosures are similarly available. The SEC provides structured JSON endpoints (data.sec.gov) that facilitate bulk downloading. The SEC's final rule document (Release 33-11216) serves as the primary labeling rubric.

https://www.sec.gov/files/rules/final/2023/33-11216.pdf
https://www.sec.gov/search-filings/edgar-application-programming-interfaces
https://data.sec.gov/

Feasibility Risks and Mitigation

The primary risk is the subjectivity of materiality judgments: the SEC rule requires disclosure of "material" incidents, but what counts as material is context-dependent. Mitigation: the classifier does not judge materiality. Instead, it classifies the content category and specificity of whatever the firm chose to disclose. Validation can be performed against known incident timelines (e.g., from breach databases such as VCDB or the Privacy Rights Clearinghouse) to test whether firms with more specific disclosures are also the ones where incidents can be independently verified. A secondary risk is the recency of the rule (effective December 2023), meaning the corpus of Item 1C filings is still growing; mitigate by beginning with the initial cohort of annual filings and expanding as new filings appear.

Expected Difficulty and Novelty

Difficulty: medium (clean data access, well-defined rubric from the SEC rule). Novelty: high (no existing validated classifier; strong regulatory demand).

Project 4: Human Capital Disclosure Specificity and Boilerplate Detection (10-K HCM)

Why It Matters

Human capital is increasingly recognized as a primary driver of firm value, yet corporate disclosure about human capital management (HCM) remains voluntary, heterogeneous, and often generic. Following the SEC's 2020 modernization of Regulation S-K Item 101, registrants are required to describe their human capital resources "to the extent material," but the rule is principles-based and does not mandate specific metrics. The result is enormous variation: some firms provide detailed workforce demographics, turnover rates, and training investment figures, while others offer only boilerplate statements. Investors, ESG analysts, and HR strategists need a tool to distinguish decision-useful HCM disclosures from generic filler at scale. Demers, Wang, and Wu (2024) provide a lexicon, cleaned data, and code for measuring human capital disclosure, and explicitly discuss fine-tuning BERT, making this a high-feasibility extension for the Ringel pipeline.

Theoretical Core

The project sits at the intersection of human capital theory (Becker, 1964)—the idea that investments in employees create intangible assets that drive productivity and competitive advantage—and disclosure theory (Verrecchia, 2001), which examines when and why firms reveal information. The SEC's own advisory committee materials (2023) note that investors increasingly request comparable HCM data, and that the principles-based approach has led to wide dispersion in disclosure quality. Demers et al. (2024) build on this by constructing a domain-specific lexicon for HCM topics (DEI, training, retention, safety, compensation, demographics, labor relations) and measuring disclosure volume and vocabulary richness across 10-K filings.

Prior Approaches and Gaps

Prior work uses keyword dictionaries, topic modeling, or simple word counts to measure HCM disclosure. Demers et al. (2024) advance the field with a curated lexicon and provide a starting point for supervised learning by pointing toward BERT fine-tuning. The remaining gap is a robust, scalable classifier that goes beyond topic detection to measure disclosure specificity—i.e., not just whether a firm mentions "training" but whether it provides decision-useful information such as training hours per employee, program names, or outcome metrics. This specificity dimension is what separates boilerplate from actionable disclosure.

Construct Definition and Label Schema

The unit of analysis is the paragraph within the HCM discussion section of 10-K filings. Two label dimensions:

Dimension 1: HCM topic. Multi-label classification, as a paragraph may address multiple topics: (a) Diversity, equity, and inclusion (DEI); (b) Training and development; (c) Employee

retention and turnover; (d) Workplace safety and health; (e) Labor relations and culture; (f) Compensation and benefits; (g) Workforce demographics and headcount; (h) None/other.

Dimension 2: Disclosure specificity. Ordinal 4-point scale: (1) Generic boilerplate—could apply to any firm; (2) Qualitative but firm-referenced—mentions programs by name without metrics; (3) Specific with metrics—includes quantitative data (e.g., turnover rate, training hours, diversity percentages); (4) Benchmarked/audited—references external benchmarks, third-party audits, or year-over-year comparisons.

Label Schema Summary

Dimension	Classes / Labels	Granularity
HCM Topic	DEI / Training / Retention / Safety / Labor relations / Compensation / Demographics / None	Multi-label per paragraph
Specificity	Generic boilerplate / Qualitative firm-referenced / Specific with metrics / Benchmarked-audited	Ordinal per paragraph

Data Sources

Form 10-K filings from SEC EDGAR, focusing on the human capital discussion in Item 1 (Business). The Demers, Wang, and Wu (2024) resources include a lexicon, data, and code published alongside their JIS article, with an arXiv preprint providing additional technical detail. EDGAR APIs support bulk downloading and full-text search. The SEC advisory committee's 2023 draft recommendation on HCM disclosure provides additional context on expected disclosure content.

https://publications.aaahq.org/jis/article/38/2/163/12398/Measuring-Corporate-Human-Capital-Disclosures
https://arxiv.org/abs/2506.10155
https://www.sec.gov/search-filings/edgar-application-programming-interfaces
https://www.sec.gov/files/20230914-draft-recommendation-regarding-hcm.pdf

Feasibility Risks and Mitigation

The main risk is industry heterogeneity: what constitutes specific HCM disclosure in manufacturing (safety incidents, OSHA metrics) differs from technology (retention bonuses, remote-work policies) or healthcare (staffing ratios, credentialing). Mitigation: use industry-stratified training and evaluation, and include industry as a feature or conditioning variable in the labeling prompt. A secondary risk is that some firms embed HCM information outside of Item 1 (e.g., in the proxy statement or a separate sustainability report); mitigate by focusing on 10-K text as the primary scope and noting the boundary condition.

Expected Difficulty and Novelty

Difficulty: medium (clean data access, strong existing resources). Novelty: moderate (specificity dimension is the key contribution).

Project 5: Exploration vs. Exploitation Strategic Orientation in Corporate Narratives

Why It Matters

The tension between exploration (searching for new knowledge, technologies, and markets) and exploitation (refining and leveraging existing capabilities) is one of the most consequential strategic trade-offs a firm faces. March (1991) established that organizations that explore too little become trapped in suboptimal competencies, while those that exploit too little fail to capture returns from their investments. A valid text-based measure of exploration–exploitation orientation would enable large-scale longitudinal studies of innovation strategy, organizational ambidexterity, and the drivers of strategic change. However, Ugur, Alturki, and Companys (2024), in a paper aptly titled "The Long March," demonstrate through extensive validity tests that the widely used dictionary-based text indicators of exploration and exploitation are likely invalid—the word lists are noisy, context-insensitive, and do not discriminate between the constructs as theorized. This creates a high-value measurement gap that an LLM-based classifier, built on a carefully designed rubric, can fill.

Theoretical Core

March (1991) defines exploration as the pursuit of new knowledge through search, variation, experimentation, and discovery, and exploitation as the refinement of existing knowledge through selection, efficiency, implementation, and execution. The two activities compete for scarce organizational resources and attention, creating a fundamental tension. Subsequent work on organizational ambidexterity (Benner & Tushman, 2003; O'Reilly & Tushman, 2013) argues that firms must manage both simultaneously, often through structural separation or contextual mechanisms. The construct has been applied across strategy, organizational behavior, innovation management, and technology management, making it one of the most broadly adopted theoretical lenses in management.

The measurement challenge is that exploration and exploitation are latent constructs that manifest through language: firms signal their strategic orientation through narratives about R&D, new product launches, process improvement, cost optimization, partnerships, and market entry. A valid classifier must distinguish between, for example, "We are investing in next-generation battery technology" (exploration) and "We are investing in manufacturing efficiency for our existing battery line" (exploitation), even though both use the word "investing."

Prior Approaches and Gaps

The dominant approach in management research is computer-aided text analysis (CATA) using word lists derived from March's (1991) original definitions. Researchers count occurrences of words like "search," "variation," "experiment" (exploration) or "efficiency," "refinement," "execution" (exploitation) in annual reports or patents. Ugur et al. (2024) subject these measures to a battery of construct validity tests—including content validity, convergent/discriminant validity, and predictive validity—and find that they largely fail. Keywords are used in unrelated contexts, the measures do not correlate with independent indicators of

exploration/exploitation, and they do not predict theoretically expected outcomes. The gap is a measurement tool that understands context, distinguishes between constructs when similar words are used, and can handle the nuance of strategic narratives.

Construct Definition and Label Schema

The unit of analysis is the sentence or short paragraph within shareholder letters, 10-K MD&A sections, and earnings-call prepared remarks. The primary classification is a four-class single-label scheme:

Label	Definition	Anchor Example (Abbreviated)
Exploration	Text describes search for new knowledge, technologies, markets, or capabilities not yet in the firm's repertoire.	We launched a pilot program to test autonomous delivery vehicles in three new metropolitan markets.
Exploitation	Text describes refinement, scaling, or optimization of existing capabilities, products, or processes.	We continued to drive efficiency gains in our core manufacturing operations, reducing unit costs by 8%.
Ambidextrous	Text explicitly discusses balancing or integrating both exploration and exploitation activities.	While investing in next-generation products, we maintained disciplined cost management in our legacy portfolio.
Neither	Text does not address strategic orientation toward novelty or efficiency (e.g., legal boilerplate, financial summaries).	The following discussion should be read in conjunction with our consolidated financial statements.

An optional enrichment layer classifies the innovation domain (product innovation, process innovation, business model innovation, market entry) to capture what is being explored or exploited.

Data Sources

Shareholder letters (often available on corporate websites or via annual-report collections), 10-K narratives (EDGAR APIs), and earnings-call prepared remarks (vendor access required for transcripts; see Project 2 notes). For initial training and validation, the publicly accessible 10-K MD&A sections provide a large, clean text source.

https://strategy.sjsu.edu/www.stable/pdf/March%2C%20J.%20G.%20%281991%29.%20Organization%20Science%202%281%29%2071-87.pdf

https://journals.sagepub.com/doi/pdf/10.1177/14761270241231724

https://www.sec.gov/search-filings/edgar-application-programming-interfaces

Feasibility Risks and Mitigation

The primary risk is construct ambiguity: some strategic initiatives simultaneously involve novelty and efficiency (e.g., "We are implementing AI to automate our existing supply chain"—is this exploration of AI or exploitation of the supply chain?). Mitigation: (1) build the labeling rubric with explicit "contrast sets" showing the same words used in different construct contexts; (2) include the "ambidextrous" class to capture genuine dual-orientation passages rather than forcing a

binary; (3) use multi-model consensus labeling (Ma et al., 2026) to surface and adjudicate disagreement cases; and (4) conduct a dedicated human-audit round focused on boundary cases. A secondary risk is that shareholder letters may reflect aspirational rhetoric rather than actual strategy; validate against independent indicators such as R&D intensity, patent filings, and new-product announcements.

Expected Difficulty and Novelty

Difficulty: medium-high (construct complexity requires careful rubric design). Novelty: high (published evidence that existing measures are invalid creates strong motivation).

Project 6: Price Fairness and Complaint Justice in Consumer Complaints (CFPB Narratives)

Why It Matters

Consumer complaints are among the highest-volume, most decision-relevant text sources available to firms and regulators. The Consumer Financial Protection Bureau (CFPB) alone publishes hundreds of thousands of consumer complaint narratives annually, each describing a specific grievance about a financial product or service. For firms, these narratives contain actionable intelligence for product redesign, service recovery, and compliance. For regulators, they signal systemic issues in pricing, lending, and servicing practices. Yet most complaint-analysis systems classify complaints by product or issue code, missing the underlying reason for consumer dissatisfaction: perceived unfairness. A classifier that detects why consumers feel treated unfairly—and what type of justice violation they describe—would transform complaint analytics from routing tools into strategic insight engines.

Theoretical Core

The project draws on two complementary theoretical frameworks. First, Xia, Monroe, and Cox (2004) provide a comprehensive conceptual framework of price fairness perceptions, defining price fairness as a consumer's assessment of whether the difference (or lack of difference) between a seller's price and a comparative standard is reasonable, acceptable, or justifiable. They identify antecedents including price knowledge, perceived seller motive, and social comparison, and distinguish between distributional and procedural dimensions of fairness.

Second, justice theory from the service-recovery literature distinguishes three dimensions of complaint justice: (a) distributive justice—whether the outcome (refund, compensation, correction) was fair; (b) procedural justice—whether the process of handling the complaint was fair (timely, accessible, transparent); and (c) interactional justice—whether the firm's representatives treated the consumer with respect, empathy, and honesty (Tax, Brown, & Chandrashekaran, 1998). Together, these frameworks provide a theoretically grounded, multi-dimensional label space for classifying consumer complaint text.

Prior Approaches and Gaps

Existing complaint classification systems (including the CFPB's own taxonomy) categorize complaints by product type (mortgage, credit card, student loan), issue type (billing dispute, account management, collections), and company response (closed with/without relief). These codes describe what the complaint is about but not why the consumer perceives unfairness or what aspect of justice was violated. The gap is a construct-aligned classifier that extracts the psychological and behavioral dimensions of consumer grievances—enabling, for example, the discovery that a spike in "procedural justice" violations at a specific firm predicts regulatory enforcement action, or that "interactional justice" failures correlate with social-media escalation.

Construct Definition and Label Schema

The unit of analysis is the paragraph or complaint segment within CFPB consumer narratives. The label schema has two primary dimensions:

Dimension 1: Unfairness type. Multi-label, as a complaint may describe multiple forms of unfairness: (a) Hidden or unexpected fee; (b) Overdraft or penalty charge perceived as excessive; (c) Interest rate or price discrimination perception; (d) Deceptive pricing or bait-and-switch; (e) Unauthorized charge or billing error; (f) Loan servicing or modification unfairness; (g) Collection practice unfairness; (h) None/other.

Dimension 2: Justice dimension violated. Multi-label across three categories: (a) Distributive justice—the outcome was unfair (e.g., "They refused to refund the charge"); (b) Procedural justice—the process was unfair (e.g., "I was transferred five times and no one could explain the charge"); (c) Interactional justice—the treatment was unfair (e.g., "The representative was dismissive and condescending").

Dimension 3: Severity. Ordinal (low/medium/high), reflecting the financial impact and emotional intensity described in the complaint.

Label Schema Summary

Dimension	Classes / Labels	Granularity
Unfairness Type	Hidden fee / Overdraft / Rate discrimination / Deceptive pricing / Unauthorized charge / Servicing / Collections / None	Multi-label per paragraph
Justice Violation	Distributive / Procedural / Interactional	Multi-label per paragraph
Severity	Low / Medium / High	Ordinal per paragraph

Data Sources

The CFPB Consumer Complaint Database is fully public and provides consumer narratives (published after the company's response or after 15 days), product/issue codes, company responses, and metadata. An API is available for programmatic access. The database is updated regularly and contains millions of records. Published narratives have already undergone PII scrubbing by the CFPB before release.

https://www.consumerfinance.gov/data-research/consumer-complaints/
https://cfpb.github.io/api/ccdb/
https://www.jstor.org/stable/30162012

Feasibility Risks and Mitigation

The primary risk is residual PII in complaint narratives, even after CFPB scrubbing (e.g., indirect identifiers, account details mentioned in context). Mitigation: apply secondary PII scrubbing using named-entity recognition and pattern matching before training. A secondary risk is that complaint text is often informal, repetitive, and emotionally charged, which may reduce label

consistency. Mitigation: use detailed anchor examples in the labeling rubric, oversample edge cases for human audit, and evaluate inter-rater agreement specifically on high-emotion texts. The human audit should also check for demographic inference risks—ensuring the classifier does not inadvertently learn to associate unfairness types with demographic groups.

Expected Difficulty and Novelty

Difficulty: medium (clean data access, well-established theory). Novelty: moderate to high (justice-dimension classification is novel in this context).

Project 7: Greenwashing Signal Detection in ESG Reports and Public Media

Why It Matters

Greenwashing—the practice of misleading stakeholders about a firm's environmental practices or the environmental benefits of its products—is among the most urgent concerns in sustainable finance, ESG investing, and corporate governance. Regulators worldwide are tightening enforcement; the EU's Green Claims Directive, the SEC's proposed climate disclosure rules, and national advertising standards bodies are all targeting greenwashing. Investors suffer from information asymmetry: they cannot easily distinguish firms with genuine environmental commitments from those engaged in symbolic impression management. A validated, scalable classifier that detects greenwashing signals in corporate text would serve investors, regulators, auditors, and academic researchers studying the boundaries between legitimate sustainability communication and deception.

Theoretical Core

The construct is grounded in legitimacy theory (Suchman, 1995), which explains that organizations manage perceptions to align with societal expectations, and that discrepancies between symbolic actions and substantive behavior generate legitimacy risk. Delmas and Burbano (2011) provide the leading management framework for greenwashing, identifying institutional, organizational, and individual drivers that lead firms to communicate misleadingly about their environmental performance. They define greenwashing as the intersection of poor environmental performance with positive environmental communication—a definition that maps naturally to a "gap detection" task between what firms say and what external evidence reveals.

Calamai, Balalau, Le Guenedal, and Suchanek (2025) provide a comprehensive survey of greenwashing detection in text, cataloging existing methods, datasets, and open challenges. They emphasize the scarcity of tailored annotated datasets and the need for methods that can distinguish between types of green claims (aspirational, qualitative, quantified, verified) and assess the evidentiary basis for each claim. The SwissText shared tasks (2023) operationalize this as a paired task: comparing ESG report language to public media coverage to detect inconsistencies.

Prior Approaches and Gaps

Prior approaches include keyword-based ESG scoring (which cannot distinguish substantive from symbolic claims), sentiment analysis of sustainability reports (which captures tone but not evidentiary quality), and ESG rating comparisons (which capture disagreement between raters but not the textual signals that drive it). The SwissText shared task provides an initial dataset pairing ESG reports with public media, but the field lacks a general-purpose classifier that can be applied to any sustainability report and that produces interpretable, multi-dimensional scores. Calamai et al. (2025) explicitly identify this as the central open challenge.

Construct Definition and Label Schema

To maintain tractability for the Ringel pipeline, the full label schema is decomposed into two independent classification tasks, each suitable for GenAI labeling and fine-tuning:

Task A: Claim strength classification. Each sentence or paragraph containing an environmental/social claim is classified on a 4-class ordinal scale: (1) Aspirational—forward-looking commitment with no specific plan or timeline (e.g., "We aspire to achieve carbon neutrality"); (2) Qualitative—describes an action or initiative without quantification (e.g., "We have implemented energy-saving measures"); (3) Quantified—includes specific metrics, targets, or data (e.g., "We reduced Scope 1 emissions by 18% year-over-year"); (4) Verified—references third-party audit, certification, or regulatory compliance (e.g., "Our Scope 1 and 2 emissions were verified by Deloitte under ISAE 3410").

Task B: Evidence presence classification. For each claim, a binary or 3-class label: (a) No supporting evidence in the text; (b) Partial evidence—some data or examples provided but incomplete; (c) Substantive evidence—claim is accompanied by specific data, methodology, or external reference.

A third, optional enrichment labels the ESG pillar (E/S/G) of each claim for downstream analysis.

Label Schema Summary

Dimension	Classes / Labels	Granularity
Claim Strength	Aspirational / Qualitative / Quantified / Verified	Ordinal per sentence/paragraph
Evidence Presence	No evidence / Partial / Substantive	3-class per claim
ESG Pillar (optional)	Environmental / Social / Governance	Single-label per claim

Data Sources

The SwissText greenwashing shared-task dataset provides paired ESG report text and public media coverage, annotated for inconsistency signals. Sustainability report corpora are available from SustainabilityReports.com, the GRI database, and individual company websites. Calamai et al.'s (2025) survey catalogs additional datasets. For public media pairing, news APIs and corporate press release archives can be used. Also, there is ESG content on the LSEG Workspace at KFBS.

https://swisstext.org/archive/2023/shared-task-1-detecting-greenwashing-signals-through-a-comparison-of-esg-reports-and-public-media/
https://www.swisstext.org/wp-content/uploads/2023/09/Greenwashing.pdf
https://arxiv.org/abs/2502.07541
https://www.sustainabilityreports.com/

Feasibility Risks and Mitigation

The primary risk is the complexity of the full greenwashing construct: a complete assessment would require comparing claims to external evidence (actions, emissions data, controversies), which goes beyond a single classification task. Mitigation: deliberately scope the classifier to detect claim strength and evidence presence within the text itself, not to render a verdict on whether greenwashing has occurred. Frame the output as "risk of misleadingness" rather than "greenwashing detected." This is both scientifically more defensible and legally prudent. For external validation, correlate classifier outputs with ESG rating divergences, environmental controversies (from RepRisk or similar databases), and regulatory enforcement actions. A secondary risk is that sustainability reports are long and heterogeneous; mitigate with pre-processing to extract claim-bearing sentences before classification and evaluate with document-level aggregation.

Expected Difficulty and Novelty

Difficulty: medium-high (construct complexity, multi-task design). Novelty: high (no existing validated, multi-dimensional greenwashing classifier; strong regulatory and investor demand).

Summary Comparison of Seven Projects

The table below summarizes the key characteristics of each project across feasibility, novelty, and impact dimensions.

Project	Primary Data	Label Type	Difficulty	Novelty	Impact
Consumer Emotions (LX)	Reviews + survey	Multi-label + ordinal	Medium	Moderate–High	Very High
Evasiveness	Earnings-call Q&A	3-class ordinal	Medium	Moderate	High
Cyber Disclosure	10-K Item 1C, 8-K	Multi-class + ordinal	Medium	High	High
HCM Specificity	10-K HCM section	Multi-label + ordinal	Medium	Moderate	High
Explore vs. Exploit	10-K, letters	4-class single-label	Med–High	High	Very High
Complaint Justice	CFPB narratives	Multi-label + ordinal	Medium	Mod–High	High
Greenwashing	ESG reports + media	Two-task ordinal	Med–High	High	Very High

Project Choice

Each team must choose one of these seven projects. Consider the trade-off between choosing a unique project or one that many other teams also chose. Many teams working on the same few projects creates direct competition in final vertical AI performance. While some teams might find this particularly exciting, others may wish to work on something more niche.

Getting Started

Make sure to look at my exemplary story (Class 21):
Smarter, cheaper, greener: Vertical AI for business analytics
http://www.ringel.ai/UNC/2026/BUSI488/Class21/Ringel_488-2026_Class21.pdf

To help you with your Capstone, I wrote a full pipeline in a python notebook that does all the key steps you need for your Capstone Project by example of classifying 10K sentences into business functions. This includes querying genAI via API at scale, creating holdout and training data sets, fine-tuning a pretrained LLM, and evaluating the performance of genAI and your fine-tuned (vertical AI) model.

What I didn't do is give you a construct of interest, collect your data, clean and preprocess your data, and draw conclusions and write reports for you.

Here is the python notebook:
http://ringel.ai/UNC/2026/helpers/Ringel_2026_VerticalAI_Capstone_Pipeline_Example.ipynb

I also created a zip file with the outputs from the above notebook (excluding the actual trained vertical AI because it is 1.5GB). All subfolders, datasets, etc. are there. This is a great blueprint for what data you need to deliver with your capstone on a shared drive (provide link to me) or uploaded if sufficiently small (less than 20MB):
http://ringel.ai/UNC/2026/helpers/Ringel_2026_VerticalAI_Capstone_Pipeline_Example.zip

The contents of the zip file also help you see what the expected output its (by example of a multi-label classification problem). You will need to adapt this code to your problem. Use genAI (e.g., Claude Opus 4.6) for this. The pipeline gives you a solid base to work off.

References

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Benner, M. J., & Tushman, M. L. (2003). Exploitation, exploration, and process management: The productivity dilemma revisited. *Academy of Management Review*, 28(2), 238–256.

Calamai, T., Balalau, O., Le Guenedal, T., & Suchanek, F. M. (2025). Corporate greenwashing detection in text—A survey. *arXiv:2502.07541*. https://arxiv.org/abs/2502.07541

Coombs, W. T. (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate Reputation Review*, 10(3), 163–176.

Cortina, L. M., & Areguin, M. A. (2021). Putting people down and pushing them out: Sexual harassment in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, 8, 285–309.

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, 54(1), 64–87.

Demers, E., Wang, Y., & Wu, J. (2024). Measuring corporate human capital disclosures. *Journal of Information Systems*, 38(2), 163–191.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, 51(3), 227–242.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1), 109–128.

Gordon, L. A., Loeb, M. P., & Sohail, T. (2010). Market value of voluntary disclosures concerning information security. *MIS Quarterly*, 34(3), 567–594.

Hollander, S., Pronk, M., & Roelofsen, E. (2010). Does silence speak? An empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research*, 48(3), 531–563.

Kiefer, L., et al. (2025). ESG-Activities: A benchmark for ESG activity detection in financial text. *arXiv:2502.21112*. https://arxiv.org/abs/2502.21112

Koch, S., & Pasch, S. (2022). CultureBERT: Measuring corporate culture with transformer-based language models. *arXiv:2212.00509*. https://arxiv.org/abs/2212.00509

Laverty, K. J. (1996). Economic “short-termism”: The debate, the unresolved issues, and the implications for management practice and research. *Academy of Management Review*, 21(3), 825–860.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, 107(1), 34–47.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Ludwig, S., Danaher, P. J., Yang, X., Lin, Y.-T., Abedin, E., Grewal, D., & Du, L. (2026). Extracting consumer insight from text: A large language model approach to emotion and evaluation measurement. *arXiv:2602.15312*. https://arxiv.org/abs/2602.15312

Ma, S., Lin, Y., & Yang, Y. (2026). EvasionBench: Detecting evasive answers in financial Q&A via multi-model consensus and LLM-as-judge. *arXiv:2601.09142*. https://arxiv.org/abs/2601.09142

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71–87.

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*, 58(3), 20–38.

O'Reilly, C. A., & Tushman, M. L. (2013). Organizational ambidexterity: Past, present, and future. *Academy of Management Perspectives*, 27(4), 324–338.

Quinn, R. E., & Rohrbaugh, J. (1983). A spatial model of effectiveness criteria: Toward a competing values approach to organizational analysis. *Management Science*, 29(3), 363–377.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12), 46–54.

Richins, M. L. (1997). Measuring emotions in the consumption experience. *Journal of Consumer Research*, 24(2), 127–146.

Ringel, D. M. (2023). Creating synthetic experts with generative artificial intelligence. *SSRN Working Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4542949

Roseman, I. J. (1991). Appraisal determinants of discrete emotions. *Cognition and Emotion*, 5(3), 161–200.

SEC (Securities and Exchange Commission). (2023). Cybersecurity risk management, strategy, governance, and incident disclosure (Release 33-11216). https://www.sec.gov/files/rules/final/2023/33-11216.pdf

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4), 813–838.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20(3), 571–610.

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing*, 62(2), 60–76.

Ugur, M., Alturki, T., & Companys, Y. E. (2024). The long march: A critical look at computer-aided text analysis measures of exploration and exploitation. *Creativity and Innovation Management*. https://journals.sagepub.com/doi/pdf/10.1177/14761270241231724

Verrecchia, R. E. (2001). Essays on disclosure. *Journal of Accounting and Economics*, 32(1–3), 97–180.

Von Solms, B., & Von Solms, R. (2004). The 10 deadly sins of information security management. *Computers & Security*, 23(5), 371–376.

Xia, L., Monroe, K. B., & Cox, J. L. (2004). The price is unfair! A conceptual framework of price fairness perceptions. *Journal of Marketing*, 68(4), 1–15.