WorldQuant Quantitative Research Internship Summary

Pranav Khetarpal | DeepHuman Algo Lab (DeepResearch) | Mumbai | Sep 29, 2025 - Dec 31, 2025

Executive snapshot

I joined WorldQuant as a Quantitative Research Intern to prototype and operationalize LLM/NLP-driven approaches for building tradable signals from text and alternative data. Across the internship, I delivered (1) an LLM-scored, directed relationship graph from SEC filings, (2) an extension of the same framework to OCR'd sell-side analyst reports, and (3) issuer-level daily features from UPC-level shipment data with robust weighting and data-quality handling. Senior researchers performed backtests on the delivered datasets; early internal tests discussed with me indicated ~1.5-1.6 Sharpe from filings, ~2.2-2.3 from analyst reports, and ~1.3-1.4 from UPC signals in a long/short market-neutral framework (high-level indication, not a final production-ready evaluation). At the end of the internship, I was encouraged to reapply; a full-time return offer was not extended and was communicated as a business/headcount decision.

Scope and constraints

- Primary mandate: use LLMs (not classical embedding similarity alone) to extract and quantify inter-company relationships from text at scale.
- Data sources: SEC 10-K/10-Q filings and OCR'd research analyst reports from major banks; plus UPC-level shipment data mapped to issuers.
- Constraint: cost/latency favored a single-pass LLM prompt for extraction + scoring (two-stage prompting was considered but deprioritized for cost).

Project 1 - SEC filings relationship graph (directed, weighted edges)

Goal: convert 10-K/10-Q narratives into a backtest-ready directed graph where each edge (A->B and B->A) captures relationship type and intensity, plus directional "trading sentiment" inferred from the text.

- Document focus: after manual review and pilot tests, concentrated on MD&A and Notes to Financial Statements as the highest-yield sections for meaningful relationships.
- Edge unit: per document, extracted the unique company-pair relationships (A,B) supported by the document context; produced PIT-aligned rows with identifiers and provenance fields for traceability.

Relationship scoring ontology (designed from scratch)

I designed an 11-factor rubric to translate textual relationship information into numerical weights. Scores were returned by the LLM as discrete intensities (0-3) or NA (not mentioned / not applicable). Ten factors were directional (scored from each firm's perspective to encode asymmetry), and one factor captured direction-invariant dynamics.

- Directional factors (examples): relationship criticality; financial materiality; regulatory opportunity; regulatory impact severity; litigation/direct financial liability; market & reputational impact; competitive & technological pressure; financial-health dependency (plus additional factors of similar intent and granularity).
- Direction-invariant factor: relationship dynamics (e.g., steady-state vs one-off vs rapidly changing relationship strength).
- Directional sentiment: for each pair, produced AB and BA "trading sentiment" (scaled experiments included -3..3 and -100..100) to represent expected market impact for one firm based only on information about the other.

Implementation details (LLM prompting + entity resolution + outputs)

- Prompting: constrained schema outputs; temperature=0 with fixed seed; iterative prompt/spec refinements based on pilot runs (including runs where per-score rationales were collected for QC and rubric calibration).
- Entity resolution: extracted company names/tickers from text (when present) and mapped to BBG identifiers using a large reference dataset and regex/fuzzy matching to handle name variants.
- Output format: parquet/CSV tables with A/B names + BBG IDs, relationship type, directional factor scores (A->B and B->A), directional sentiment (AB/BA), point-in-time (PIT), source identifiers, and a short overall rationale used for qualitative QC.

Scale, yield, and observed failure modes

- Coverage: typical yield of ~2-5 meaningful relationships per filing; most extracted entities mapped successfully to instrument identifiers.
- Scale: ran the pipeline over 2020-2024 filings (on the order of ~12k filings).
- Key failure mode: under high cognitive load (many tasks + long context in a single prompt on a small, cost-efficient model), outputs sometimes collapsed to default/near-zero scores across features; mitigated via prompt iterations, but a stronger next iteration would add gating and selective re-asks.

Project 2 - Analyst-report relationship graph (OCR'd sell-side reports)

After completing the filings graph, I extended the same relationship extraction + scoring framework to OCR'd sell-side analyst reports to cover a second core text corpus used in quant research.

- Pre-processing: removed OCR artifacts/special characters; stripped boilerplate sections (e.g., disclaimers); removed tables/figures to focus on narrative text.
- Scoring: reused the relationship taxonomy and scoring rubric; adapted scoring scale in later iterations (e.g., 0-100) while maintaining directional structure.
- Scale and yield: processed ~400k reports (2020-2024) with typical yield of ~1-3 meaningful relationships per report; delivered structured edge tables for downstream backtesting.

Project 3 - UPC shipment alternative data (issuer-level daily features)

Goal: build robust issuer-level daily aggregates and features from UPC-level shipment data (price, cartons/volume, origin factory hash, destination, etc.) and reduce noise from long-tail UPCs via adaptive weighting/gating.

- Aggregation & feature families: daily sales/volume per issuer; active UPC counts; concentration/entropy across origins, destinations, and UPC mix; WoW/MoM/YoY growth for UPCs and issuers; additional QC/health features to handle missingness and structural breaks.
- Time-varying UPC weighting (three central schemes): (1) economic value share over last 4 quarters combined with stability (1 - coefficient of variation); (2) consistency-weighted (week-to-week stability); (3) co-movement weight (correlation between UPC growth and issuer growth excluding that UPC).
- Binary gating variants: converted continuous weights into percentile thresholds (e.g., 70/80/90/95th) to include only the most informative UPCs in aggregates.
- Data-quality handling: addressed NaNs, vendor change artifacts around 2020, returns, and discontinuations using custom rules/features; used Polars/lazy execution for scale where helpful.

Initiated (deprioritized) - Macro risk exposure from filings

Planned second-phase work was to quantify firm-level macro risk exposures from filings using an LLM-driven taxonomy and scoring rubric. Work began with taxonomy definition and early prompt/schema exploration, but the project was deprioritized in favor of completing analyst-report relationships and the UPC stream.

Deliverables and handoff

- Datasets: PIT-aligned edge tables for filings and analyst reports; issuer-level daily UPC aggregates and feature tables (parquet/CSV).
- Research artifacts: relationship taxonomy, 11-factor rubric definitions, prompt/spec iterations, QC rationale format, and a set of ablation/testing ideas for senior QRs to explore.
- Engineering artifacts: clean, reproducible pipelines with parallelized CPU execution and tracking fields for stability and lineage.

Note: This summary is written to be non-proprietary and reproducible; it describes the work at a level suitable for external review while avoiding firm-internal details. Backtest numbers above are reported as the initial internal indications communicated to me and are not a guarantee of production performance.