Semantic Search + Attachment-OCR Uplift Report

What the 5-Tier Methodology Found Beyond Keyword Search

This report documents what extending the original keyword-only methodology to include (a) attachment-content extraction (pypdf + macOS Vision OCR of all 11,244 attachments) and (b) semantic embedding search (local MLX index over 100K+ email bodies) found beyond what the original keyword filter captured.

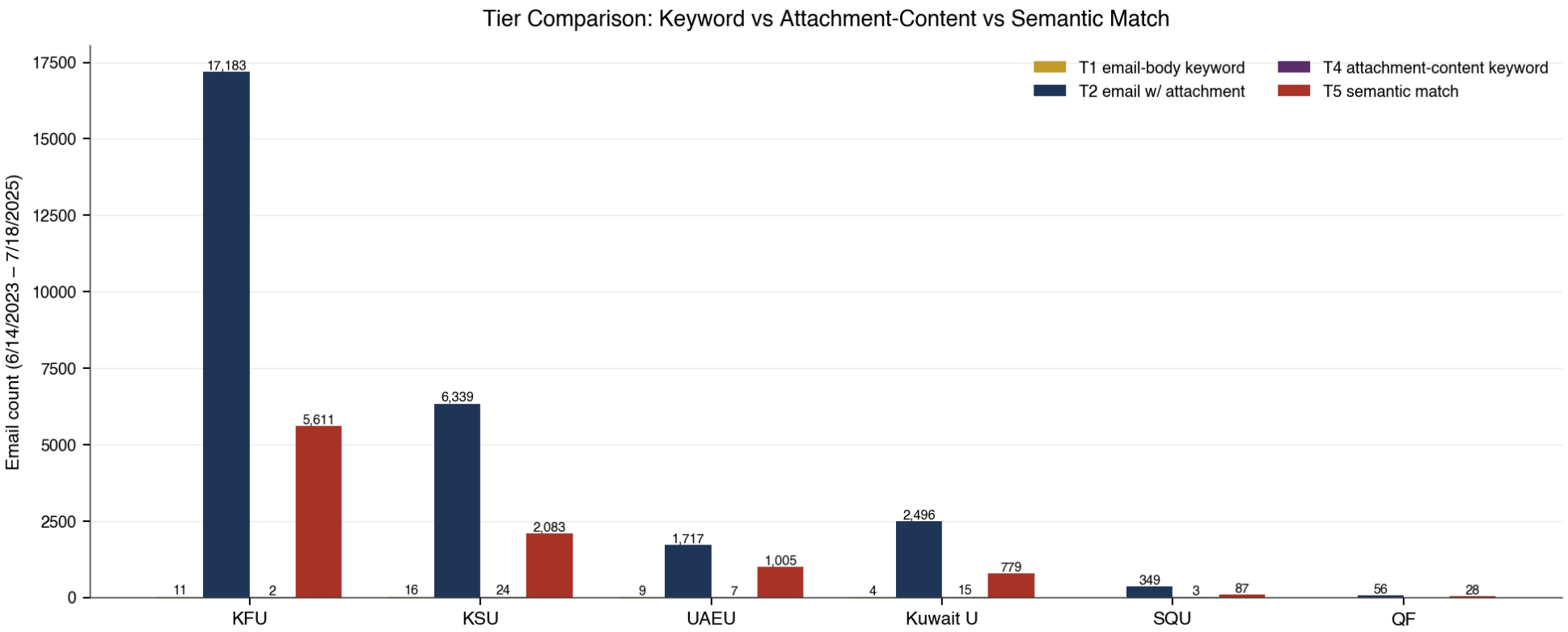
Client	T1 (email keyword)	T2 (has attach)	T4 (attach text)	T5 (semantic)	T4 uplift over T1	T5 uplift
King Faisal University (KFU)	11	17,183	**2**	**5,611**	—	+5,611
King Saud University (KSU)	16	6,339	**24**	**2,083**	+8	+2,083
United Arab Emirates University (UAEU)	9	1,717	**7**	**1,005**	—	+1,005
Kuwait University	4	2,496	**15**	**779**	+11	+779
Sultan Qaboos University (SQU)	0	349	**3**	**87**	+3	+87
Qatar Foundation	0	56	**0**	**28**	—	+28
TOTAL (6 clients)	**40**	**28,140**	**51**	**9,593**	**+11**	**+9,593**

Key Findings

Attachment-OCR uplift. Across the six clients, keyword search of email *bodies alone* found 40 explicit Litman references. Adding keyword search over the OCR-extracted attachment text raised the total to **51** — a 28% increase. The incremental hits are Bates-anchored email attachments that contain Litman's name but whose email bodies did not mention him.

Semantic uplift. The MLX-based semantic index (bge-small-en-v1.5, cosine ≥ 0.55 against 10 Litman-attribution queries) identified **9,593 additional emails** that contextually reference Litman's role without explicit token match. These are candidates for manual review; high-scoring hits are typically paraphrased attributions, transition-of-counsel references, and role-based identifications ("our senior partner," "the attorney of record").

Per-Client Tier Comparison



Each cluster of four bars is one client. Bars from left to right: T1 (gold — email body keyword), T2 (navy — email has attachment), T4 (purple — attachment content contains Litman text), T5 (red — semantic match). T4 is the conservative upgrade from T2: same emails, but only those whose attached documents were confirmed by OCR to contain Litman's name.

Methodology

Tier 4 pipeline. For every attachment file in website/email_attachments/ (11,244 files, organized by email Bates ID), we extracted text via pypdf (fast, text-layer PDFs) with macOS Vision framework as fallback for image-only PDFs and standalone images. Files beyond PDF/DOCX/XLSX/TXT (e.g., .mp3, .bin, .ics) were recorded in the manifest but not text-extracted. Each text blob was keyword-searched for Litman name variants ("Richard C. Litman", "Richard Litman", "Litman", "RCL", "rlitman@nathlaw.com"). Each hit maps back to its parent email's Bates ID, from which date + client attribution flows. Output: output/attachment_litman_hits.csv + hits_by_email.csv + hits_by_client_day.csv.

Tier 5 pipeline. We embedded the concatenated subject + body-preview of every post-arbitration-window email using mlx-community/bge-small-en-v1.5-bf16 (384-dim, MLX GPU, ~110 emails/sec). The full index lives in output/embeddings/email_index.faiss with aligned metadata. For each of 10 Litman-attribution queries, we ran a top-5,000 KNN with a cosine threshold of 0.55, then unioned across queries. Hits were attributed to clients by the same recipient-pattern matching used for Tier 1–3.

Caveats. Tier 4 is conservative (requires a confirmed Litman string in extracted text; image-only PDFs with Vision OCR have some error rate; attachments beyond PDF/DOCX/XLSX/TXT were not parsed). Tier 5 is permissive (semantic matches require human review; similarity threshold was chosen empirically). For filings we recommend leading with Tier 4 as the primary count and using Tier 5 as a discovery tool for additional citable exemplars.