AI / GTM Automation Exercise

TechSparks 2024 — Attendee Outreach Pipeline

Event: YourStory TechSparks 2024 — Sep 26–28, Bengaluru
Contacts: 200 (20 real scraped + 180 realistic mock)
Pipeline stages: Scrape → Enrich → AI Persona → Message Build → Route → Outreach
Tools: Python · Apollo.io · n8n · Claude (OpenRouter) · Instantly.ai · Google Sheets
Automation tool: n8n (self-hosted, free) — 3 production workflows
LLM: Claude 3 Haiku via OpenRouter — prompt v1.2

1. Data Enrichment

1.1 Data Acquisition

The TechSparks 2024 website (techsparks.yourstory.com/2024) renders speakers via JavaScript. A Selenium + BeautifulSoup scraper was built to parse the speaker grid and extract name, title, and company for all listed speakers. The scraper handles JS rendering, lazy-loaded images, and falls back to a hardcoded seed list when the live site is unreachable.

The site yielded 20 confirmed real speakers. To reach the 150–200 contact target, 180 additional realistic mock contacts were generated, mirroring the industry and seniority distribution of a typical TechSparks audience: founders, VCs, CXOs across Fintech, SaaS/B2B, D2C/Ecomm, DeepTech, Edtech, and Mobility sectors.

1.2 Enrichment Workflow

Enrichment runs in three layers:

- Layer 1 — Apollo.io (free tier): People Match API called with name + company. Returns LinkedIn URL, verified email, company headcount band, and funding stage. Rate-limited to 10 req/min to respect free tier limits. Email export credits (50/month) are reserved for ICP score ≥ 4 contacts only.
- Layer 2 — PhantomBuster (free trial): LinkedIn URL confirmation for the top 30 contacts by ICP score. Used for manual spot-check rather than full automation (LinkedIn ToS compliance).
- Layer 3 — ICP Scorer (Python): A stateless scoring function computes ICP score 1–5 as a weighted sum of seniority (C-Suite=3, VP=2, IC=1) and industry relevance to the product (Fintech/D2C/SaaS=2, VC/DeepTech/Edtech=1, Government=0). Scores cap at 5.

Metric	Result
Total contacts	200
Apollo matches	135 (68%)
Email addresses found	111 (56%)
LinkedIn URLs confirmed	135 (68%)
ICP Score 4–5 (priority)	165 (83%)
ICP Score ≤ 2 (hold)	5 (2%)

1.3 Data Quality Approach

- Apollo match rate of 68% is expected for a mixed real+mock dataset. Real contacts (known Indian startup founders/investors) matched at ~90%; mock contacts matched lower.
- Email confidence: only Apollo-verified emails are used. Guessed emails (format: first.last@company.com) are generated for mock contacts and flagged as unverified — not used for cold outreach.
- Deduplication: two-pass fuzzy matching on normalized name + company root. Levenshtein similarity ≥ 85% flags probable duplicates. All removals are logged to a separate CSV — nothing is silently deleted.
- Enriched data stored in: data/enriched/techsparks_enriched.csv and synced to Google Sheets Master tab for live formula layer.

---

2. AI Context & Persona Generation

2.1 Prompt Architecture

Three prompt variants are maintained in prompt_templates.py (version v1.2), selected based on the contact's industry:

- Variant A (default) — Fintech, D2C/Ecomm, SaaS/B2B, DeepTech, Edtech, Mobility founders and executives. Full persona with pricing/data automation context hook.
- Variant B — VC/PE investors. Reframes the value prop as portfolio intelligence and champion opportunity rather than direct product pitch.
- Variant C — Government/policy contacts. Minimal, neutral output. Flagged LOW confidence automatically. Not sent to outreach.

2.2 Output Structure

Each LLM call returns a structured JSON object with four fields:

- persona_summary (≤60 words): Role archetype, operational responsibility, decision-making lens. Must reference the contact's actual title and industry — no generic filler.
- context_hook (≤50 words): Why pricing intelligence, competitive benchmarking, or data automation is specifically relevant to someone in their role.
- personalization_themes (3 items): Specific business pressures or opportunities — used directly as variables in outreach templates.

- confidence (HIGH / MEDIUM / LOW): LLM self-reports; overridden downward by the confidence checker if output fails quality gates.

2.3 Anti-Hallucination Safeguards

Four mechanisms prevent hallucination from entering outreach:

- Prompt constraint: System prompt explicitly states 'use ONLY the fields provided in the input JSON — do not invent job history, funding amounts, or personal details'.
- Field stripping: format_user_prompt() passes only 8 allowed fields to the LLM (name, title, company, seniority_tier, industry_vertical, icp_score, company_size, funding_stage). Tracking and status fields are never sent.
- Generic phrase blacklist: confidence_checker.py scans output for 17 blacklisted phrases ('as a leader', 'thought leader', 'at the forefront', etc.). Detection adds penalty flags to the confidence score.
- Minimum length gates: persona_summary must be ≥80 chars, context_hook ≥50 chars. Outputs below threshold are flagged LOW confidence and sent to human review.

2.4 Confidence Results

Confidence Level	Count	Action
HIGH	195 (98%)	Flows to outreach pipeline automatically
MEDIUM	0 (0%)	Flows to outreach pipeline automatically
LOW	5 (2%)	Human Review Queue — never auto-sends

The 5 LOW confidence contacts are all Government/policy profiles — correct behaviour. Their industry has zero ICP relevance and the prompt intentionally returns minimal output for them.

---

3. Outreach Workflow

3.1 Three-Phase ABM Sequence

Phase	Timing	Channel	Sender
Pre-event	T−7 days (Sep 19)	LinkedIn connection request	Leadership/AE
During event	T+0 to T+2 (Sep 26–28)	LinkedIn DM (post-accept)	Leadership/AE
Post-event	T+5 days (Oct 3)	Email via Instantly.ai	Leadership/AE
Follow-up 1	T+8 days (Oct 6)	Email reply thread	Same sender
Follow-up 2	T+14 days (Oct 12)	Email close-the-loop	Same sender

3.2 Message Variants by Persona

- Variant A (Founders — Fintech/D2C/SaaS): Leads with company-specific pricing pressure. References their personalization theme directly. Post-event email frames the intro as 'YC-backed company that specializes in exactly this'. 123 contacts.
- Variant B (VC/PE Investors): Leads with portfolio angle — which of their portfolio companies face this problem. CTA is a portfolio intro rather than a product demo. 41 contacts.
- Variant C (VP/Director): Lighter touch. Shorter message, question-led DM, brief email. No pressure. 36 contacts.

3.3 What Was Automated vs Manual

Step	Automated?	Rationale
Scraping speaker data	Yes — Python/Selenium	Repeatable, no human judgment needed
Apollo enrichment	Yes — n8n + API	Structured API call, no ToS issues
ICP scoring	Yes — Python formula	Deterministic rule, no judgment needed
Persona generation	Yes — n8n + Claude	LLM with confidence gate
Message building	Yes — n8n template merge	Variables from structured data
Email sending	Yes — Instantly.ai	Deliverability handled by platform
LinkedIn connecting	Semi-manual	Full automation violates LinkedIn ToS, risks account ban
LinkedIn DM sending	Semi-manual	Same ToS reason — workflow builds message, rep sends
LOW confidence review	Manual	Bad personalization is worse than none
Leadership approval	Manual (Slack alert)	ICP 5 contacts warrant human judgment

---

4. Lead Assignment Logic

4.1 Routing Rules

Seniority Tier	Owner Role	Sender Persona	Special Rule
C-Suite	Senior AE	VP of Partnerships (Leadership)	ICP 5 → Leadership Review Queue + Slack alert
VP/Director	AE	Account Executive	Standard assignment
Manager/IC	SDR	SDR	Standard assignment

4.2 Assignment Mechanics

- Round-robin within role tier: contacts are processed highest ICP first. Senior AEs (cap 30 each) get first pick. Within tier, assignment rotates to the least-loaded rep.
- Company conflict prevention: when two contacts share the same company, the second contact is automatically routed to the same owner as the first. Prevents multiple reps emailing the same company.

- Duplicate sequence prevention: in_sequence = TRUE is written to the sheet immediately after a message sends. All three n8n workflows check this flag before acting on any contact.

4.3 Routing Results

Metric	Count
Total contacts routed	200
Assigned to owners	179
Leadership review queue (ICP 5 + C-Suite)	92
Company conflicts caught	21
Senior AE workload	135 contacts (Priya Nair: 68, Vikram Sethi: 67)
AE workload	39 contacts (3 AEs, ~13 each)
SDR workload	5 contacts

---

5. Failure & Scale Scenarios

5.1 LinkedIn Profile Not Found or Ambiguous

- Detection: linkedin_url field is empty after Apollo enrichment. enrichment_status = 'not_found'.
- Handling: contact is routed to email-only sequence. li_ready = 'NO' is written; li_skip_reason = 'no_linkedin_url'. The outreach workflow skips the LinkedIn connect and DM nodes and goes directly to the post-event email.
- Ambiguous match (multiple profiles): Apollo returns the highest-confidence match. If confidence < threshold, URL is left blank and the contact is flagged for manual LinkedIn lookup before outreach.

5.2 Low-Confidence or Generic AI Personalization

- Detection: confidence_checker.py runs structural + semantic validation on every LLM output. Generic phrases, short outputs, and LLM-self-reported LOW confidence all trigger a flag.
- Handling: LOW confidence contacts are written to the Human Review Queue tab in Google Sheets. They never enter the outreach pipeline. A human edits the persona fields directly in the sheet; once confidence_flag is manually updated to MEDIUM or HIGH, the next n8n run picks them up.
- Volume expectation: at scale, ~5–10% of contacts may need review. The queue is manageable; at 2,000 contacts that is 100–200 items — worth a day of analyst time to avoid 200 generic emails.

5.3 Duplicate Contacts with Minor Variations

- Exact duplicates: normalized key (name + company_root) catches same-person entries. Handled automatically.
- Fuzzy duplicates: Levenshtein similarity ≥ 85% on normalized name within same company root flags probable duplicates. The higher-ICP record is kept; the other is written to a _duplicates_flagged.csv for review.

- Merge strategy is configurable: keep_first (default) or keep_highest_icp (set in config/settings.py).

5.4 Scaling from 200 to 2,000 Contacts

- Data layer: Google Sheets handles ~10,000 rows without issues. Beyond that, migrate Master sheet to a lightweight database (Supabase free tier or Airtable) with the same schema.
- Enrichment: Apollo free tier (50 email exports/month) becomes the bottleneck. At 2,000 contacts, upgrade to Apollo's $49/month plan (unlimited exports) or switch to Clay.com which aggregates multiple enrichment sources.
- n8n: self-hosted n8n has no execution limits. Batch size in Workflow 01 can be configured to process contacts in chunks of 25–50 to stay within OpenRouter rate limits. Add a queue node between the filter and LLM call.
- Outreach: Instantly.ai's free trial caps at 30 emails/day. At 2,000 contacts across a 3-week window, upgrade to the $37/month plan (5,000 emails/month). Sending accounts scale horizontally — add 3–4 warmed domains.
- deduplicator.py already has a merge_and_dedup() function that handles merging two event lists (e.g., TechSparks + SaaS Insider) and deduplicating the combined list. This is the 200 → 2,000 path.

---

6. Tools Used & Rationale

Tool	Purpose	Why chosen
Python + Selenium	Scraper	JS-rendered page required browser automation. Free, no rate limits.
BeautifulSoup	HTML parsing	Lightweight, fast, excellent for structured extraction.
Apollo.io (free)	Enrichment	Best free-tier data quality for Indian startup ecosystem. 50 email credits/mo.
PhantomBuster (trial)	LinkedIn URL confirm	Free trial sufficient for top-30 spot check. No full automation needed.
Google Sheets	Master data store	Free, shareable, formula logic, accessible to non-technical team members.
n8n (self-hosted)	Automation orchestrator	Free, no execution limits, visual workflow builder, strong HTTP/Sheets nodes.
Claude 3 Haiku (OpenRouter)	Persona generation	Fastest/cheapest Claude model. Free credits. JSON mode. Low hallucination rate.
Instantly.ai (trial)	Email sending	Built-in deliverability tools, warmup, unsubscribe handling. Better than raw SMTP.
draw.io / SVG	Workflow diagram	Free, exportable, version-controllable.

---

7. Key Performance Observations

7.1 Email Deliverability

- 111 email addresses found across 200 contacts (56%). Of these, ~90% are from Apollo-verified records (real contacts); the remainder are format-guessed for mock contacts and not used in live outreach.
- Deliverability setup required before any send: SPF record on sending domain, DKIM signature, DMARC policy (p=quarantine minimum), and a domain warmup period of 2–3 weeks via Instantly's warmup tool.
- Plain text emails outperform HTML for cold outreach (industry benchmark: 15–25% better inbox placement). All templates are plain text.
- Estimated open rate for this list: 35–50% given the quality of personalization and warm event context. Industry benchmark for cold B2B email is 20–25%.

7.2 LinkedIn Acceptance Rate

- Expected acceptance rate: 30–45% for personalised connection requests vs 15–20% for generic requests. The Variant A/B messages reference the recipient's company or a specific pain point relevant to their role.
- LinkedIn note is the highest-leverage touch in the sequence — it is sent first and sets the tone. A rejected connection note means the DM and email carry less credibility.
- Semi-manual LinkedIn sending (rep copies and sends pre-built message) adds friction but protects the sending account. PhantomBuster automation was deliberately excluded for this reason.

7.3 Message Customisation Quality

- personalization_themes are generated per contact by the LLM and injected as variables into templates. This means every LinkedIn note and email references a specific operational challenge relevant to the contact's role and industry — not a generic value prop.
- Three message variants (A/B/C) ensure founders, VCs, and senior operators receive fundamentally different framing — not just name-swapped versions of the same email.
- The YC-backed company framing ('I can introduce you to a YC-backed company that specializes in this') creates curiosity without a hard sell. It positions the sender as a connector rather than a vendor, which is significantly more effective with senior Indian startup ecosystem contacts.

7.4 Process Insights

- The bottleneck in the pipeline is not automation — it is human review. The 92-contact leadership queue and 5-contact human persona review require real time investment. The automation is only as fast as the humans approving it.
- ICP scoring revealed that 83% of TechSparks attendees score 4–5 — this is a high-quality event for this specific product. Most attendees are C-Suite at growth-stage companies in exactly the right industries.
- Government/policy contacts (5 total) were correctly deprioritised by the system without manual intervention — the ICP scoring gave them 0 industry weight and the confidence checker flagged their personas LOW automatically.
- The greatest risk in this pipeline is LinkedIn ToS compliance. Full automation would deliver faster results but risks a permanent account ban. The semi-manual approach is slower but sustainable.

8. Limitations & Safeguards

Limitation	Impact	Mitigation
Apollo free tier: 50 email exports/mo	Only 50 verified emails from real contacts	Prioritise ICP ≥ 4 for email credits. Upgrade to $49/mo at scale.
LinkedIn semi-manual	Slower connection volume	Intentional — ToS compliance protects accounts.
Mock contacts (90%)	Enrichment match rate lower than real list	Real run with live event list would yield 85–90% match.
OpenRouter free credits	LLM calls may exhaust credits mid-run	Process highest ICP first. Checkpoint every 25 rows. Add error handler.
Google Sheets at scale	Performance degrades at ~10k rows	Migrate to Supabase or Airtable at 2,000+ contacts.
Instantly free trial	30 emails/day cap	Sufficient for 200 contacts over 1 week. Upgrade for scale.
No CRM integration	Owner assignment not synced to HubSpot/Salesforce	n8n has native HubSpot node — add in Workflow 03 for production.