
Benchmarking the Personalization Capabilities of Large Language Models

Ashutosh Srivastava Siddharth Yedlapati Vinay Aggarwal Yaman K Singla
Shashwat Dixit Jitendra Ajmera Balaji Krishnamurthy



Adobe Media & Data Science Research

behavior-in-the-wild@googlegroups.com

Abstract

Personalization, the act of varying a message to induce action from a specific receiver while keeping sender, channel, and time fixed, has a long tradition in psychology and marketing as a two-party problem in which sender and receiver have independent objectives. Large language models remove the bounded-inventory constraint of classical retrieval-and-ranking approaches by generating a continuum of message variants conditioned on inferred receiver state, raising the question of how well current models perform personalization in the classical sense. Existing LLM personalization benchmarks measure sender-side adaptation, in which the receiver is the same user the model is serving. The two-party question, whether a generated message induces its intended action in a third party, has been investigated only through A/B tests and small-scale human studies that cannot be re-run against a new model on demand. We adapt the Bayesian Persuasion framework of Kamenica and Gentzkow (2011) to generative agents and instantiate the formulation in sales, where receiver actions are routinely logged against the outreach that induced them. We release **SDR-Bench**, a public corpus of 6,279 customer success stories spanning 22 industries and approximately 200 enterprises, served through a temporally constrained simulation that prevents future-data leakage. Across frontier LLMs and deep-research agents, we observe a consistent personalization plateau and on a Fortune 100 tech cohort no model statistically separates successful from unsuccessful outreach. A field deployment with 12 professional sales representatives validates the framework, with 48 percent of model-generated content rated immediately useful and senior-expert agreement at Pearson 0.82. We release **SDR-Arena** and **SDR-Bench** publicly to support reproducible study of generative personalization at scale.

1 Introduction

Communication as defined by the seminal work of Lasswell (1) characterizes any communicative act in five variables: who says what to whom, through which channel, with what effect, and at what time. Within this framework personalization can be located as the act of varying the message, that is, the *what*, while conditioning on (and holding fixed) the speaker, receiver, channel, and time. The act has two parties whose interests need not initially align; a sender, who selects the message with the goal of inducing some action from the receiver, and a receiver, who has independent preferences and chooses whether to act. The study of how senders shape messages to induce receiver action has been a long tradition in multiple fields, including, psychology, economics, and marketing, beginning with the Yale Communication and Attitude Change program (2) and continuing through the Elaboration Likelihood Model (3) and the formal signaling-game treatment of Bayesian Persuasion (4).

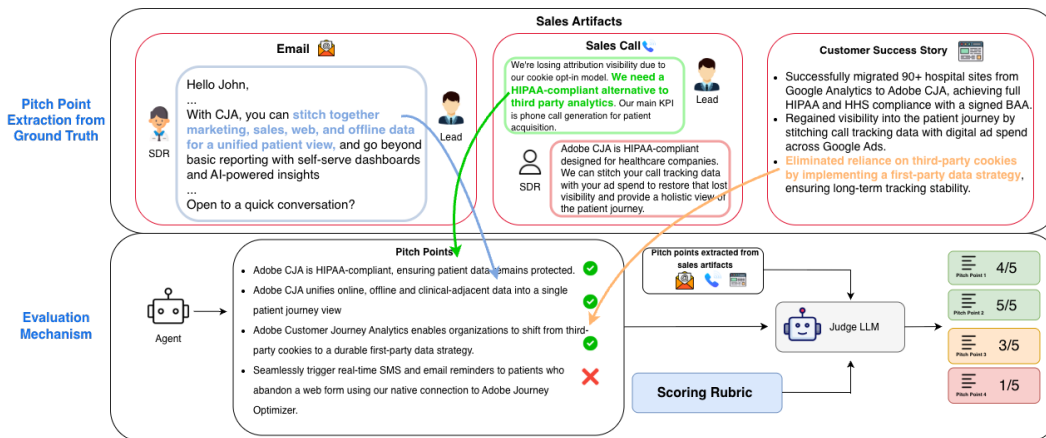


Figure 1: Overview of SDR-Arena showcasing how LLM generated output is compared with artifacts like Sales Emails, Transcripts & Success Stories to benchmark their personalization capability

So far, machine learning research on personalization has approached this problem as learning a policy to retrieve and rank items from a fixed inventory of candidates. Recommender systems rank items from a catalog against user-history signals (5); advertising platforms optimize the selection, and targeting of ads among a fixed pool of pre-authored creatives (6); persona-based dialogue systems condition response generation on explicit persona representations while remaining restricted to comparatively narrow conversational domains (7).

Large language models transition this personalization from a retrieval and ranking problem to a generative one. LLMs can generate a continuum of message variants for a given (speaker, receiver, channel, time) tuple, conditioned on whatever attributes of the receiver can be inferred from the available context. A growing body of work has applied LLMs to generative-personalization in marketing (8), education (9; 10), and human-AI interaction (11). These applied results raise the question of how well current LLMs perform personalization in the sense the classical literature studies it: as a sender selecting *what* to say in order to induce a specific receiver action. While some work exists for measuring LLM personalization, however, it measures a very different property compared to the personalization talked about in psychology and economics literature (1; 4). LaMP (12) evaluates personalized text generation conditioned on a user’s history of past interactions. PersoBench (13) measures persona consistency in open-domain dialogue. PersonaConvBench (14) scores persona-grounded conversational quality. PersonaLens (15) evaluates assistant behavior under declared user preferences. PersonaMem (16) measures long-horizon recall of user attributes across sessions. In these works, the recipient of the LLM generated message is the same user that the model is serving, and the optimization target is one-party: the alignment between the model’s output and the preferences of the user who issued the prompt, in the same sense that RLHF aligns an assistant to its user. The two-party question, in which sender and receiver have independent objectives and the sender’s success is measured by *whether the receiver acts*, has primarily been investigated through randomized human-subject experiments in which participants are exposed to human- or LLM-generated persuasive messages and evaluated based on subsequent shifts in attitudes, agreement, or behavioral intentions(8; 17). Such studies are tied to a specific methodology, require weeks of execution, and cannot be re-run against a new model on demand. Consequently, comparisons between LLMs and human experts for two-party personalization remain specific to individual studies and are not directly comparable across systems. Therefore, there is a need for a formal model of personalization which accounts for both the sender and receiver of the message, sufficient to support a reproducible and automated benchmark applicable to arbitrary generative systems.

This formulation requires an empirical setting where receiver actions are observed and logged against the specific messages that induced them, the messages are authored at the receiver level rather than at the segment level, and human-authored ground truth at known successful induced actions are available at scale. Sales outreach artifacts provides a good testbed to measure this as receiver actions (replying, scheduling a call, closing a deal) are routinely logged against the specific outreach that induced them (18). A sales outreach is drafted one-to-one by a Sales Development Representative (SDR) for a specific prospect. In an in-house study conducted with a Fortune 100 enterprise, we observed that

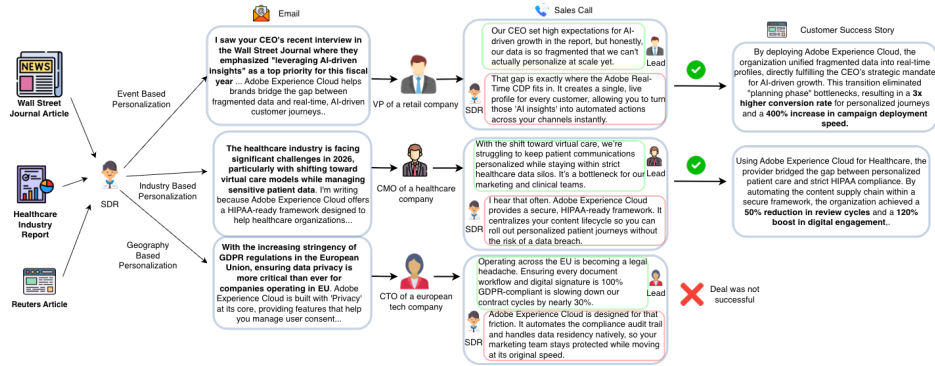


Figure 2: Sales Journey: From Prospecting to Outreach to Call and eventual Deal Closure leading to Success Story publication

personalized SDR outreach achieved approximately seven times the click-through rate of templated outreach when promoting the same products to comparable prospect groups.. Furthermore, the sales funnel produces a layered set of human-authored artifacts at known successful transitions, namely outreach emails that secured a call, call transcripts that secured deal discussions, and post-deal customer success stories that document the content that closed the deal. Together, these properties make sales a natural empirical instantiation of the two-party formulation where each artifact in the funnel serves as ground truth at a different stage of the same (seller, prospect, product) tuple, enabling stage-specific evaluation of generative personalization.

We develop our framework **SDR-Arena** (illustrated in Fig 1) on this empirical setting and list our contributions as follows:

- We adapt Bayesian Persuasion to generative agents, recovering personalization as informational alignment between an agent’s generated content and the receiver-specific content implicit in the ground-truth sales outreach artifact.
- We construct SDR-Bench, a public corpus of 6,279 customer success stories from approximately 200 enterprises across 22 industries, each paired with the seller, prospect, product, and historical timestamp required to evaluate whether an agent can predict the strategic content of the deal-closing pitch.
- We release SDR-Arena, an evaluation framework that operationalizes the formalization on SDR-Bench and on proprietary sales artifacts; to prevent future data leakage, where an agent retrieves the very success story it is being asked to predict, SDR-Arena serves agents a frozen view of the public web at the historical timestamp of each evaluation instance.
- We apply the framework to proprietary sales-email and sales-transcript corpora from a Fortune 100 tech company and a mid-sized healthcare firm, comprising approximately 115,000 filtered outreach emails and 5435 outreach calls by 124 SDRs labeled by whether they induced a successful receiver action.
- We validate the framework through field deployment with 12 professional sales development representatives across our partner enterprises and a gold-standard exercise with senior SDRs from five enterprises

Across frontier LLMs and open-source deep-research agents, including STORM (19), ODR(20), GPT-4o(21), and Qwen-2.5(22), we observe a consistent personalization plateau. Alignment scores cluster in the 30 to 43 percent range, and on the tech-firm cohort no model statistically separates successful from unsuccessful outreach. Specialized agents such as STORM reach the upper end of the range, but at one to two orders of magnitude greater inference cost; standard LLMs with temporally constrained search occupy a more compute-efficient frontier.

2 Problem Formulation

We formalize the empirical sales in the form of generative personalization by first describing the Sales Development Lifecycle, then casting personalization as a Bayesian Persuasion task where the generated outreach is aimed at inducing specific actions from the recipients within the sales funnel.

2.1 Sales Development Lifecycle

A sales journey (illustrated in Fig. 2) begins with Sales Development Representatives (SDRs) researching prospective accounts to identify needs and budget signals, then sending tailored outreach emails to schedule an initial call. The call further develops the prospect’s needs and progresses toward deal closure, with some opportunities materializing into deals and others not. Following a successful closure, the workflow often culminates in a **Customer Success Story**: a publicly documented case study published by the Seller company showcasing how their products helped a customer overcome key challenges. Released as web articles, these stories validate the partnership by documenting the transition from a ‘pain state’ to a ‘success state’ (23). Examples include success stories from Oracle, Salesforce, and Adobe.

2.2 Bayesian Persuasion Formulation

We adapt the Bayesian Persuasion framework of Kamenica (4) to model personalized outreach as a signaling game between a Sender (the SDR or the LLM agent that replaces the SDR) and a Receiver (the prospect). This framework naturally aligns for personalization because it acknowledges that the receiver enters the interaction with a prior belief about their own needs, and the sender’s role is to provide a signal, the personalized message, that updates the receiver’s posterior in favor of action.

The unobserved receiver state ω represents the latent compatibility between the receiver’s requirements and the sender’s product. We decompose ω at time t as a tuple $\omega_t = \{n_i, w_i\}$ where n denotes the receiver’s explicit needs (functional requirements, current pain points) and w denotes their latent wants (strategic goals, avenues for value generation). The receiver chooses an action $a \in \{0, 1\}$, where $a = 1$ represents a successful transition to the next stage of the sales funnel (e.g., an email leading to a call, or a call leading to deal discussions), while $a = 0$ represents a failure to progress.

Following Kamenica (4), the Receiver is treated as a rational Bayesian agent with a prior belief μ over ω . The Receiver takes action $a = 1$ if and only if their expected utility u_R , conditional on their belief, exceeds a reservation threshold τ :

$$\mathbb{E}[u_R(a = 1, \omega_t, \xi) \mid \mu] \geq \tau$$

The Sender’s utility u_S is aligned with the Receiver acting $a = 1$. Thus, the Sender’s goal is to deliver a signal that updates the receiver’s posterior belief such that the above condition is satisfied. Note that the utility function is also subject to exogeneous factors ξ (timing, organizational urgency, prior context, noise) that are independent of the sender’s signal but contribute to the receiver’s utility. The sender’s signal can shift μ in favor of acting; it cannot control ξ . Consequently, even an optimal signal is not guaranteed to induce $a = 1$, and the receiver’s action is best understood as a probabilistic outcome whose likelihood the sender attempts to maximize.

The key adaptation for our setting is that the sender does not observe ω_t directly. The sender (in our case an LLM agent Φ) operates on an observable context W_t (state of the world at time t) consisting of public information available at time t , from which the latent state must be inferred. From this inference, the agent generates an outreach O , which we represent structurally as a list of pitch points. Each pitch point is a specific argument linking the seller’s product to one of the receiver’s inferred needs or wants through a particular value proposition. Formally, the agent’s policy is a mapping:

$$O = \{pp_1, pp_2, \dots, pp_k\}, \hat{O} = \{\hat{p}p_1, \dots, \hat{p}p_n\} = \Phi(W_t, P \mid \hat{\omega})$$

where $\hat{\omega}$ is the agent’s inferred receiver state, \hat{O} is the outreach generated by agent Φ and P is the product. The generated list serves as the informational signal intended to update the receiver’s posterior such that the probability of the desired stage transition ($a = 1$) is maximized.

2.3 Evaluation Methodology and Proxy

The ideal objective of the Sender is to generate an outreach O that maximizes the Receiver’s expected utility: $O^* = \arg \max_O \mathbb{E}[u_R(a = 1, \omega) \mid O]$. Direct evaluation of this objective is intractable as the receiver’s utility function u_R and the true state ω are unobservable to the agent (and to the benchmark). We therefore replace the unobservable utility with an observable proxy - the informational overlap between the agent’s generated content and the content implicit in a human-authored message that is known to have induced the desired action.

We utilize a dataset of successful historical outreach attempts across different engagement stages (Emails, Calls, and Success Stories). Let $\mathcal{D} = \{(C_i, \omega_i, O_i^*)\}_{i=1}^N$ be a set of ground truth examples where O_i^* is a human-authored message that successfully induced action $a = 1$ (moving to the next funnel stage). Success in email is defined by an outreach O_i^* that led to a call; a call success implies O_i^* led to deal discussions; and a success story implies O_i^* led to deal closure. Because each O_i^* resulted in a positive outcome, it empirically satisfies the receiver’s utility threshold and can be treated as a sample from the set of utility-maximizing messages for the corresponding (sender, receiver, product, time) tuple. We extract a list of ground truth pitch points V^* from O_i^* and define the Weighted Coverage Score (WCS) as the semantic alignment between the predicted pitch points \hat{O} and the ground truth points V^* :

$$\text{Weighted Coverage Score} = \mathcal{S}(\hat{O}, V^*)$$

where $\mathcal{S}(\cdot, \cdot)$ is a semantic alignment scoring function (defined in Section 3.1).

This allows us to rigorously benchmark Agent performance by measuring the **Relevance Alignment** between the predicted pitch points (where the value proposition is embedded) against those extracted from the successful ground truth. This serves as a tractable proxy for the Bayesian persuasion objective: a higher matching score implies the Agent has successfully identified the winning strategy that induces the desired action.

WCS is, by construction, a *lower bound* on personalization quality. It captures the *what* component of personalization — whether the agent has correctly inferred the receiver’s decision-relevant needs and wants and identified the value-generating arguments that historically induced the desired action while abstracting away the *how* (style, tone, formatting). An agent that achieves high WCS has demonstrated that it can recover the receiver-specific strategic content of a known successful message; an agent that achieves low WCS has not, regardless of how well-written its output is.

2.4 Dataset Construction

To evaluate generative personalization in real-world settings, we construct two complementary datasets: (i) **SDR-Bench**, a large-scale public benchmark derived from enterprise sales artifacts, and (ii) a **private enterprise dataset** containing real sales outreach emails and downstream sales outcomes from two organizations. Together, these datasets provide both reproducible public evaluation and high-fidelity validation on real-world personalized communication.

SDR-Bench Dataset. To enable reproducible public benchmarking of generative personalization systems, we construct SDR-Bench, a large-scale corpus of publicly available enterprise sales narratives and customer success stories. We targeted approximately 12,000 global enterprises with revenues exceeding \$1B, identifying sitemaps for 8,298 organizations and collecting 117,000 candidate URLs using heuristics tailored to common success-story paths (e.g., /customer-stories, /case-study).

A multi-stage filtering pipeline (Table 3) removed non-text formats, generic landing pages, articles lacking verifiable publication dates or identifiable product solutions, and anonymized stories where the customer organization was not explicitly named (e.g., “a large food products company”). This process yielded a final corpus of 6,279 success-story articles spanning 22 industries. Distributions of companies and stories by industry are shown in Fig. 4 and Appendix Fig. 6. Detailed construction steps are provided in Appendix A.1.

Private Enterprise Dataset. To validate our theoretical proxy, we require settings where the ground-truth message O_i^* and its successful outcome ($a = 1$) are explicitly observed. We therefore collaborated with two enterprises—a Fortune 100 technology company and a mid-sized healthcare firm—to collect real human-authored sales outreach paired with downstream prospect actions.

For the Fortune 100 company, we collected approximately 100k outreach emails authored by 124 SDRs and 5,435 sales call transcripts over a two-year period (2023–2025), identifying 13,236 instances in which the outreach successfully induced a sales call. For the healthcare firm, we analyzed 24,506 outreach emails, of which 354 resulted in a scheduled sales call. These successful outreach instances serve as observed realizations of optimal messages (O^*) in our relevance-alignment framework. Table 1 summarizes the dataset construction pipeline.

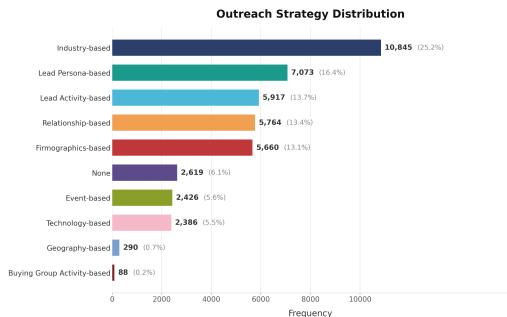


Figure 3: Distribution of count of strategies across a random subset of 34,000 emails

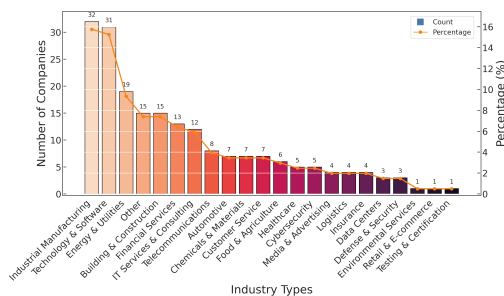


Figure 4: Distribution of count of companies by Industry Type

Human Personalization Strategies. To characterize the qualitative structure of expert personalization, we analyzed the strategies employed by SDRs across both datasets (Fig. 3). The three most common strategies were: (i) **industry-based personalization**, tailoring content to sector-specific trends and pain points; (ii) **persona-based personalization**, adapting the value proposition to the recipient’s organizational role; and (iii) **activity-based personalization**, leveraging behavioral signals such as webinar attendance or prior engagement.

Appendix Fig. 5 provides qualitative examples showing how SDRs adapt the same product positioning differently across recipients with distinct inferred needs (n_i) and wants (w_i).

Table 1: Processing of Enterprise Sales Email Data

Metric / Artifact Category	Healthcare	Tech
Number of SDRs	3	124
Total Emails Collected	48,150	609,191
Deduplication	31,034	186,379
Sales Outreach Emails	24,506	90,809
Sales Call scheduled	354	13,236
Golden dataset handpicked	400	400

3 SDR-Arena

We introduce **SDR-Arena**, a scalable framework designed to systematically benchmark LLM-based agents on generative personalization over sales outreach artifacts. To ensure a rigorous and valid evaluation, the arena utilizes an isolated environment that provides agents access to a **Historical Internet Simulator** (W_t). The arena serves as a standardized testbed for comparing diverse agentic workflows, ranging from complex research pipelines to simple tool-use configurations. We evaluate two primary configurations on SDR-Bench and our Enterprise Dataset:

- **LLMs + Web Search:** A baseline equipping frontier models with standard search tools to measure the marginal utility of agentic workflows against simpler tool-use capabilities.
- **Deep Research Agents:** Specialized agents that produce comprehensive research via multi-turn conversation and broad search retrieval over the internet(19; 20).

Historical Internet Simulator: This environment prevents “future leakage” by enforcing a strict temporal boundary, ensuring agents only synthesize information that was publicly available at the simulated time of the sales interaction. The system enforces the W_t boundary by passing `search_start_date` and `search_end_date` parameters to the BrightData SERP API (24). By restricting results to time t , we ensure that the generated pitch points are constructed solely from context that would have been accessible to a human researcher at the time of the original sales event, preventing the model from ‘cheating’, where an agent might mistakenly find the successful outcome of a deal that hasn’t happened yet in the simulation.

3.1 Evaluation Framework

We define each evaluation instance as a tuple (S, C, P, t) , where S is the seller, C is the prospect, P represents the products, and t is the historical timestamp. The tuple is extracted from each sales artifact individually. Please refer Appendix A.4 for examples.

Implementation: The agent is prompted to act as a sales representative for S pitching P to C using the time-restricted search tool. The resulting output $\hat{O} = \{\hat{p}p_1, \dots, \hat{p}p_n\}$ is a set of personalized pitch points intended to address the inferred needs and strategic goals of the prospect. We employ an LLM-based semantic judge to extract ground truth pitch points V^* from the historical sales artifact. We use the raw content of the sales artifact and employ GPT-4o (21) to perform an ontological extraction of ‘Pitch Points.’ Each pitch point is required to follow a strict **triad structure**: *Product/Service* \rightarrow *Specific Pain Point* \rightarrow *Value Proposition/Mechanism*. To ensure the pitch points are grounded, the extraction model was instructed to provide and validate pitch points with exact ‘evidence quotes’ from the source text for every claim. An expert study verified the precision and recall of this extraction to be 0.92 and 0.97 respectively, showing strong alignment with expert judgment. Refer to App. Section A.8 for more details.

For each $pp \in V^*$, the judge evaluates whether the agent’s output \hat{O} successfully covered the point. Performance is measured by the **Coverage Score**, defined as the fraction of ground truth strategic value propositions successfully recovered by the agent. We employ a **Coverage Judge** relying on a **5-point Likert scale** that grades Sales Effectiveness and Factual Precision, ranging from **0 (Miss / Irrelevant)** through **1 (Marketing Fluff)**, **2 (Topic Match)**, **3 (Implied / Soft Match)**, **4 (Strong Sales Argument)**, up to **5 (Strategic Bullseye)** - A perfect extraction that captures the exact pain point of the recipient and the specific mechanism the product provides to address it. The full scoring rubric and the judge prompt are in the Appendix.

Weighted Coverage Score (WCS): This metric normalizes the Likert-scale evaluations into a percentage representing the agent’s *completeness* in capturing the winning sales logic. For a given success story with N ground truth pitch points, let $s_i \in \{0, \dots, 5\}$ be the score assigned by the judge for the i -th point. The WCS is calculated as:

$$WCS = \left(\frac{\sum_{i=1}^N s_i}{5N} \right) \times 100\%$$

A score of 100% implies that the agent successfully predicted every critical deal-winning argument with maximum specificity. It is important to note that this is a **prediction task** rather than a retrieval task: the ground truth serves as a *future artifact*, and agents must predict these winning points using only historical data available at time t . This metric measures the semantic alignment (Section 2.3) of agent outputs against these *future artifacts* resulting in a realistic back testing scenario.

Table 2: Ground truth artifact evaluation. Values reported as Coverage.

Model	Enterprise Sales Emails					Success Stories					Average Cost per Query		
	Healthcare (<1B)		Tech (>10B)		Other	Tech.	Mfg.	Energy	IT	Agg.	Prompt	Completion	Cost
	Unsucc	Succ	Unsucc	Succ	Sales Call Transcripts	(30)	(30)	(30)	(30)	(180)	Tokens	Tokens	(\$)
STORM-QWEN-2.5	22.46	32.27	43.15	39.24	30.43	43.40	41.59	39.24	44.60	42.51	~29k	~6.3k	~0.135
ODR-QWEN-2.5	15.40	30.41	38.51	39.82	22.55	30.16	30.95	35.28	35.33	33.53	~66k	~8.6k	~0.250
QWEN2.5-72B (WEB)	32.11	36.72	39.53	36.43	25.34	32.09	36.75	37.04	38.21	36.84	~5.6k	~0.25k	~0.002
Claude Sonnet 4.6 (WEB)	59.89	60.89	64.08	61.33	34.52	52.6	60.5	54	54.9	55.8	~79.2k	~2.2k	~0.270
GPT-4o (WEB)	33.70	39.17	42.01	44.61	23.5	30.73	39.77	32.42	35.85	36.38	~7.2k	~0.3k	~0.021

4 Results & Analysis

We evaluate models across two categories: frontier LLMs augmented with the temporally-restricted SDR-Arena web-search tool, comprising **Claude Sonnet 4.6**, **GPT-4o**, and **QWEN-2.5-72B** and deep research agents, **STORM** and **ODR**, both built on QWEN-2.5-72B. These configurations are evaluated across three corpora. The first is a public corpus of curated customer success stories partitioned by industry: *Technology*, *Manufacturing*, *Energy*, and *IT*, with an aggregate set of 180 stories. The second is a corpus of transcripts of sales calls from a company exceeding \$10B in revenue. The third is a corpus of human-authored enterprise sales emails, divided into two cohorts: a *Healthcare* company with under \$1B in revenue and a *Technology* company exceeding \$10B in revenue, each containing 200 successful and 200 unsuccessful emails.

4.1 Discussion of Empirical Findings

We observe several notable trends. First, **Claude Sonnet 4.6** leads all models with an aggregate WCS of 55.8 on the public success story dataset, representing a meaningful gap above the next best model, STORM-QWEN-2.5 at 42.51, and roughly 19 points above GPT-4o. Despite this it only recovers roughly half of the strategic content of the human-authored success story, indicating a clear personalization plateau across all agent families. Second, frontier LLMs are a more cost-efficient alternative to deep research agents. Claude Sonnet 4.6 achieves the highest WCS at an inference cost comparable to ODR (\$0.270 vs. \$0.250), while surpassing it by more than 20 WCS points.

The **Enterprise Sales Email** cohorts (Table 2) reveal a sector-dependent pattern. In the *Healthcare* cohort, models more consistently assign higher scores to successful outreach than unsuccessful outreach (e.g., STORM: 32.27 vs. 22.46), suggesting they capture personalization cues relevant to specialized, high-stakes sectors. In the *Technology* cohort, however, this pattern inverts or collapses: several models score unsuccessful emails comparably to or higher than successful ones (e.g., STORM: 43.15 vs. 39.24), indicating that models generate coherent but strategically shallow content insufficient to drive real-world revenue in competitive markets.

Pre-training leakage is not driving WCS A complementary concern is that publicly indexed success stories in **SDR-Bench** may have appeared in LLM pre-training corpora, inflating WCS through memorization rather than genuine inference. To probe this, we partition SDR-Bench by article publication date and re-evaluate on pre-2024 vs. post-2024 cohorts; since GPT-4o’s training cutoff sits between Q4-2023 and early 2024, post-2024 stories are unlikely to have been seen during pre-training. We observe negligible WCS differences across the split (STORM: 0.42 vs. 0.43; GPT-4o: 0.36 vs. 0.36). The absence of pre-training-era inflation indicates that performance on **SDR-Bench** reflects context-conditioned synthesis, not retrieval of memorized content.

4.2 Human Alignment and Validation

To ensure that our automated metrics reliably reflect real-world quality, we conducted expert studies calibrating our Coverage Judge against independent human raters, and confirming practical utility through deployment with professional sales representatives.

To show that the Coverage Judge follows human judgement, we conducted a human study on 20 success stories (80 model responses across STORM, ODR, GPT, and Qwen). Three independent human annotators, blinded to model identities, scored coverage following the exact protocol of our LLM-based Coverage Judge.

The study yields three convergent signals on Judge fidelity. (i) The Judge tracks human scores with strong rank correlation (Spearman’s $\rho = 0.7435$, $p < 0.0001$), holding across models (ODR: 0.7768; GPT-4o: 0.7575; QWEN-2.5: 0.7330; STORM: 0.6963). (ii) The Judge *preserves model ordering*: both human- and Judge-graded WCS rank STORM > GPT-4o > ODR, so absolute-score differences do not distort comparative conclusions. (iii) The Judge is systematically more conservative than human raters (STORM: 48.06 vs. 55.29; GPT-4o: 37.94 vs. 46.04; ODR: 31.10 vs. 41.69), ruling out score inflation and establishing WCS as a rigorous lower bound that tracks human intuition at scale.

We show that the WCS-based ranking transfers to expert SDR judgment in two field studies with 12 senior SDRs from the partner enterprises whose data appears in this paper. The first measures *per-pitch usefulness* - whether any individual model-generated pitch point would be used verbatim in real outreach, and the second measures *strategy-level overlap* between model output and SDR-authored gold standards. Together they probe the two granularities at which an automated score can mismatch expert judgment: per-point quality and overall strategic match.

- **Per-pitch usefulness:** Twelve SDRs used GPT-4o + SDR-Arena to generate pitch points for 200+ new prospect companies inside their normal outbound pipeline, with each SDR auditing the model output on accounts they were actively working. For every generated pitch point, the SDR rated, on a binary criterion, whether it both (a) reflected genuine understanding of the prospect’s pain points and (b) was usable in outreach without rewriting; **48.2%** of pitch points met both criteria. The field rate corresponds to roughly half of agent output being expert-grade in live deployment, with the remaining points being factually accurate but strategically generic, directly consistent with the personalization plateau identified above.

- **Gold-Standard Alignment:** We asked senior SDRs (≥ 10 years of industry experience and ≥ 5 years at the firm) from five enterprises to independently author reference “gold-standard” strategies to pitch 30 products to 5 prospects each. The SDRs were not shown any model output during the exercise, so the reference strategies are an independent expert read of what *should* be pitched. We then computed the overlap between the gold-standard strategy and the outputs of the four benchmarked models, and correlated this expert-overlap score against the corresponding automated WCS on SDR-Bench. Across models, expert overlap and WCS track at Pearson $r = 0.816$. The WCS-based ranking therefore transfers to senior-SDR judgment without re-tuning the rubric per enterprise, supporting WCS as a calibrated proxy for whether an agent has identified the strategic content a domain expert would pitch.

Together, these studies establish that our pipeline’s outputs are both factually grounded and meaningful in live sales contexts.

We also evaluate a closed source deep research agent, **GEMINI-2.5-PRO-DR** on a separate 25-story subset. Its Deep Research API does not expose temporal-restriction parameters and its higher inference cost precludes broader evaluation. On this subset, it achieves a WCS of **62.63**. Notably, the margin between this score and that of Claude Sonnet 4.6 with web search remains narrow, further underscoring that frontier LLMs with web search constitute a cost-effective alternative to deep research agents.

5 Conclusion

In this work, we introduced **SDR-Arena**, the first comprehensive framework for benchmarking the generative personalization capabilities of Large Language Models. By grounding our evaluation in the **Bayesian Persuasion** framework, we transitioned from subjective assessments of “quality” to a rigorous measure of **Relevance Alignment**. Our experiments utilize **SDR-Bench**—a novel, high-fidelity corpus of over 6,200 success stories—and a unique enterprise-scale dataset of successful sales outreach to quantify how effectively LLMs can synthesize winning strategic arguments.

Our findings reveal a significant “personalization plateau,” showing a substantial gap remains between AI-generated outreach and human-level strategic proficiency.

By releasing **SDR-Arena**, we provide the research community with the tools necessary to study autonomous personalization while strictly controlling for data leakage. As LLMs continue to move into high-stakes business operations, we hope this framework serves as a foundation for developing AI agents that are not only persuasive but verifiably aligned with the nuanced needs of their human recipients.

References

- [1] H. D. Lasswell, “The structure and function of communication in society,” in *The communication of ideas*. Harper and Row, 1948, vol. 37, pp. 215–228.
- [2] C. Hovland, I. Janis, and H. Kelley, *Communication and Persuasion: Psychological Studies of Opinion Change*, ser. Yale paperbound. Yale University Press, 1953. [Online]. Available: https://books.google.co.in/books?id=ZYizW6_P-goC
- [3] R. E. Petty and J. T. Cacioppo, “The elaboration likelihood model of persuasion,” ser. *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. Academic Press, 1986, vol. 19, pp. 123–205. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0065260108602142>
- [4] E. Kamenica and M. Gentzkow, “Bayesian persuasion,” *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.
- [5] F. Ricci, L. Rokach, and B. Shapira, “Introduction to recommender systems handbook,” in *Recommender systems handbook*. Springer, 2010, pp. 1–35.
- [6] H. Choi, C. Mela, S. Balseiro, and A. Leary, “Online display advertising markets: A literature review and future directions,” *Information Systems Research*, vol. 31, 06 2020.

- [7] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, “Personalizing dialogue agents: I have a dog, do you have pets too?” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. [Online]. Available: <https://aclanthology.org/P18-1205/>
- [8] S. Matz, S. Vaid, H. Peters, G. Harari, and M. Cerf, “The potential of generative ai for personalized persuasion at scale,” *Scientific Reports*, vol. 14, 02 2024.
- [9] O. Tasdelen and D. Bodemer, “Generative ai in the classroom: effects of context-personalized learning material and tasks on motivation and performance,” *International Journal of Artificial Intelligence in Education*, pp. 1–22, 2025.
- [10] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, “The role of large language models in personalized learning: a systematic review of educational impact,” *Discover Sustainability*, vol. 6, no. 1, pp. 1–24, 2025.
- [11] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang *et al.*, “When large language models meet personalization: Perspectives of challenges and opportunities,” *World Wide Web*, vol. 27, no. 4, p. 42, 2024.
- [12] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, “Lamp: When large language models meet personalization,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 7370–7392.
- [13] S. Afzoon, Z. Jamali, U. Naseem, and A. Beheshti, “Persobench: Benchmarking personalized response generation in large language models,” *arXiv preprint arXiv:2410.03198*, 2024.
- [14] L. Li, P. Cai, R. A. Rossi, F. Deroncourt, B. Kveton, J. Wu, T. Yu, L. Song, T. Yang, Y. Qin *et al.*, “A personalized conversational benchmark: Towards simulating personalized conversations,” *arXiv preprint arXiv:2505.14106*, 2025.
- [15] Z. Zhao, C. Vania, S. Kayal, N. Khan, S. B. Cohen, and E. Yilmaz, “Personalens: A benchmark for personalization evaluation in conversational ai assistants,” in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 18 023–18 055.
- [16] B. Jiang, Y. Yuan, M. Shen, Z. Hao, Z. Xu, Z. Chen, Z. Liu, A. R. Vijjini, J. He, H. Yu *et al.*, “Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory,” *arXiv preprint arXiv:2512.06688*, 2025.
- [17] E. Durmus, L. Lovitt, A. Tamkin, S. Ritchie, J. Clark, and D. Ganguli. (2024) Measuring the persuasiveness of language models. [Online]. Available: <https://www.anthropic.com/news/measuring-model-persuasiveness>
- [18] H. Terho, A. Salonen, and M. Yrjänen, “Toward a contextualized understanding of inside sales: the role of sales development in effective lead funnel management,” *Journal of Business Industrial Marketing*, vol. 38, no. 2, pp. 337–352, 09 2022. [Online]. Available: <https://doi.org/10.1108/JBIM-12-2021-0596>
- [19] Y. Shao, Y. Jiang, T. Kanell, P. Xu, O. Khattab, and M. Lam, “Assisting in writing Wikipedia-like articles from scratch with large language models,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 6252–6278. [Online]. Available: <https://aclanthology.org/2024.naacl-long.347/>
- [20] LangChain, “Open deep research,” https://github.com/langchain-ai/open_deep_research, 2025, accessed: 2026-05-07.
- [21] OpenAI, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: <https://arxiv.org/abs/2303.08774>

- [22] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, “Qwen2.5 technical report,” *arXiv preprint arXiv:2412.15115*, 2024. [Online]. Available: <https://arxiv.org/abs/2412.15115>
- [23] H. Terho, A. Salonen, and M. Yrjänä, “Toward a contextualized understanding of inside sales: the role of sales development in effective lead funnel management,” *Journal of Business & Industrial Marketing*, vol. 38, 09 2022.
- [24] Bright Data. (2026) Bright data serp api. Accessed: 2026-05-07. [Online]. Available: <https://brightdata.com/products/serp-api>

A Appendix

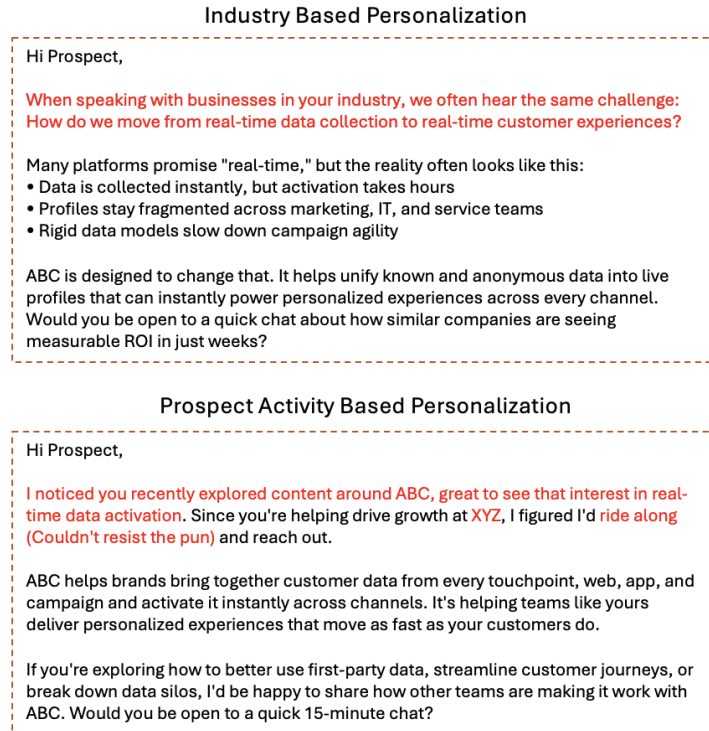


Figure 5: Personalization in actual Sales Emails

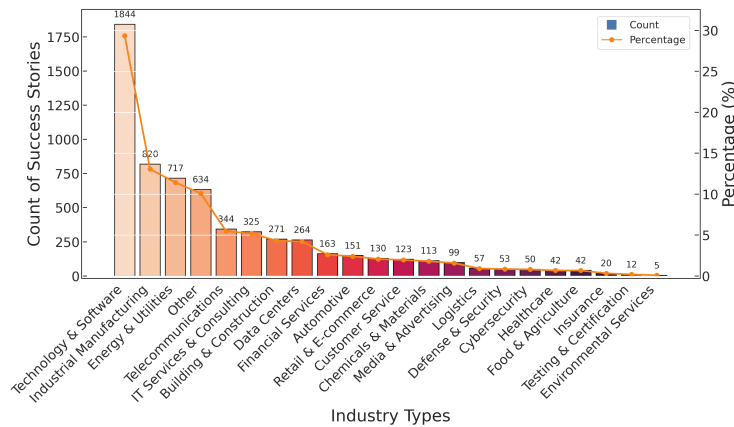


Figure 6: Distribution of count of Success Stories by Industry Type

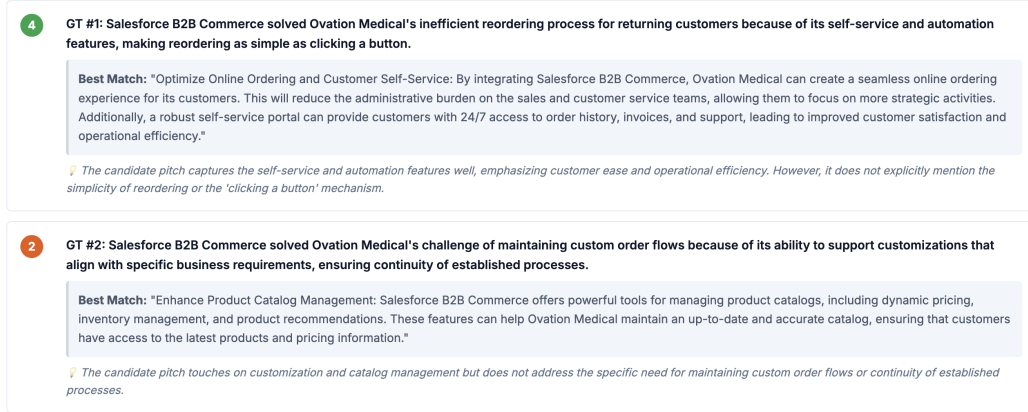


Figure 7: **Qualitative Example:** Ground truth pitch points scored against pitch points generated by the agent

A.1 SDR-Bench: Dataset Curation Details

Filtration Criteria	Count
Domains Found for Companies with over \$1B revenue	~30k
Domains Found for B2B Companies with over \$1B revenue	12,080
Companies whose Sitemap could be found	8,298
Candidate Success Story URLs based on pattern matching	~117k
Count of Companies covering these 117k URLs	1,772
Exclude non-text formats (videos/pdfs)	~79k
URLs for which content could be collected	~31k
Qwen based filtering using content to exclude listicle, parent, generic pages and pages with no publish date	~7.2k
Filtering out stories where the customer is not a specific business	6279

Table 3: Filtration Criteria and Counts for Scraped Public Data

A.2 Sales Emails

A.2.1 Filtering & Analysis

Let $\mathcal{E} = \{e_1, e_2, \dots, e_N\}$ denote the raw corpus of sales emails. We apply a three-stage filtering pipeline:

- **Language Filtering:** We remove all non-English emails using language detection, yielding $\mathcal{E}_{en} \subset \mathcal{E}$.
- **Email Deduplication:** We identify and remove duplicate email templates using a combination of exact matching and fuzzy string comparison yielding $\mathcal{E}_{deduplicated} \subset \mathcal{E}_{en}$
- **Intent Classification.** We employ an LLM-as-a-judge paradigm to classify emails into outreach versus non-outreach categories. Specifically, we filter out generic conversational emails, administrative correspondence, and non-sales communications. Let $\mathcal{J} : e \rightarrow \{0, 1\}$ be the LLM judge function where $\mathcal{J}(e) = 1$ indicates a valid sales outreach email. Our filtered corpus is thus:

$$\mathcal{E}_{filtered} = \{e \in \mathcal{E}_{deduplicated} : \mathcal{J}(e) = 1\}$$

For each email $e \in \mathcal{E}_{filtered}$, we use an LLM to extract the set of strategies employed in each email: $\text{Strat}(e) \subseteq \mathcal{S}$. This allows us to visualize the following patterns:

- **Strategy Frequency Distribution:** The distribution $P(s)$ over strategies reveals the current state of human personalization practices.

- **Product-Conditional Strategies:** The distribution $P(s|\text{Product}k)$ identifies product-specific personalization patterns.

These distributions provide interpretable insights into how human SDRs currently operationalize personalization.

Beyond strategy classification, we extract fine-grained pitch points from each email using an LLM. For each email e , we extract:

$$\mathbf{pp}(e) = \{pp_1, pp_2, \dots, pp_k\}$$

where each pp_i represents a discrete pitch point used in the outreach. These pitch points constitute the ground truth against which DR agent outputs are evaluated.

The email dataset comprises of annotated outreach emails with the following attributes per sample:

- **Target Company T_i :** The recipient organization.
- **Sender’s Company S_i :** The sender’s organization.
- **Email E_i :** The content of the email
- **Timestamp t :** Date when the email was sent
- **Product P :** The solution being pitched
- **Strategy Labels $\text{Strat}(e)$:** Personalization strategies used in the email
- **Pitch Points $pp(e)$:** Pitch points used in the email

A.2.2 Personalization Strategies

In order to systematically characterize the various personalization strategies used in the emails, we employed the following two step pipeline:

First, we asked domain experts to manually annotate a seed set of emails to identify recurring personalization patterns. **Second**, we used an LLM to extract and cluster strategies from 500 randomly sampled emails, which were then reconciled with expert annotations to produce a unified taxonomy.

A personalization strategy $s \in \mathcal{S}$ is a variable representing the primary information source leveraged to establish relevance between the seller’s value proposition and the buyer’s needs. We define the Personalization Strategy Space $\mathcal{S} = \{s_1, s_2, \dots, s_{10}\}$ consisting of 10 categories:

- **Industry based:** References industry-specific trends, pain points, competitors, or case studies from the target company’s industry.
- **Event based:** Leverages trigger events (funding rounds, MA, product launches, earnings reports, news mentions) to identify timely business needs.
- **Technology based:** References the recipient’s current tech stack to propose replacement, integration, or complementary solutions.
- **Lead Activity-based:** References direct actions by the specific lead (whitepaper downloads, webinar attendance, pricing page visits, demo interactions).
- **Buying Group Activity-based:** References collective actions by the lead’s team or buying committee.
- **Geography-based:** Utilizes physical location or regional regulatory context (e.g., GDPR, CCPA compliance requirements).
- **Lead Persona-based:** Explicitly maps the lead’s role, title, or job responsibilities to role-specific pain points.
- **Firmographics-based:** Leverages company-level metrics (headcount growth, revenue, department size) as personalization anchors.
- **Relationship-based:** References existing customer relationships, cross-sell or upsell opportunities.
- **None:** Generic outreach lacking recipient-specific context.

A.3 How to measure personalization in an ideal world?

Ideally, one could evaluate personalization by observing how the same Receiver responds to multiple personalized signals s_i , where each s_i is generated by a different LLM, effectively a multiverse of interventions. By comparing Receiver actions across these interventions, we could directly quantify the personalization abilities of different LLMs. Because such a multiverse is unavailable in practice, we construct an empirical benchmark using real-world sales interactions.

A.4 Task Formulation Details

For the success story of Salesforce, the tuple would be (S: Salesforce, C: Snapology of Lehi, P: Salesforce Starter, t: 26-05-2023).

A.5 Alignment of LLM with humans for pitch point extraction

TP	FP	FN	Precision	Recall	F1 Score
138	11	3	0.92	0.97	0.95

Table 4: Comparative analysis between LLM extracted pitch points and human annotations on 30 customer success stories

A.6 Token Usage vs Performance of Agents

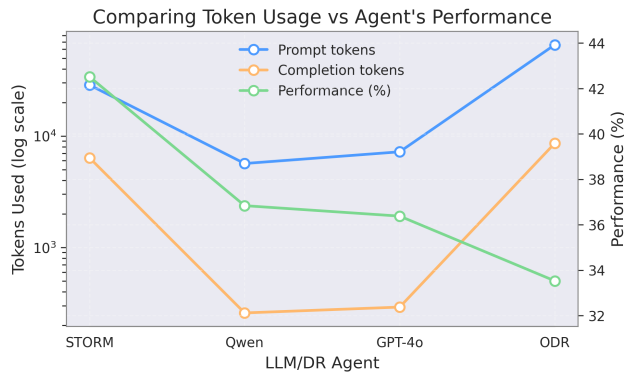


Figure 8: Graph of token usage vs performance of various agents

Table 5 reports the average per-outreach token consumption and inference cost for each agent configuration on the SDR-Bench evaluation set, alongside its WCS. Costs are computed using public list prices for the corresponding model API at the time of evaluation. Deep-research pipelines (STORM, ODR) consume one to two orders of magnitude more tokens than standard LLM-plus-search baselines, while only marginally improving WCS over the latter. QWEN-2.5-72B is the most cost-efficient configuration, achieving WCS within ~ 5.7 points of STORM at $\sim 67\times$ lower cost.

Table 5: Per-outreach inference cost vs. WCS on the SDR-Bench evaluation set.

Model	Avg. Prompt Tokens	Avg. Completion Tokens	Avg. Inference Cost	WCS
STORM	~29k	~6.3k	~\$0.135	42.51
QWEN-2.5-72B	~5.6k	~250	~\$0.002	36.84
GPT-4o	~7.2k	~300	~\$0.021	36.38
ODR	~66k	~8.6k	~\$0.250	33.53
Claude Sonnet-4.6	~79k	~2.0k	~\$0.270	33.13

A.7 Prompts Library

LLM as a Judge Evaluation Prompt

SYSTEM:

Role: You are a Senior Sales Enablement Evaluator.

Your goal is to determine if an AI agent has successfully extracted the "Winning Pitch Points" from a customer success story.

INPUT DATA:

1. **GROUND TRUTH (GT) WINNING POINTS:**

(These are the proven, specific facts that won the deal)

<<gt_pitch_points_str>>

2. **CANDIDATE PITCH (Predicted):**

(These are the points generated by the AI agent)

<<candidate_pitch_points_str>>

TASK:

For EACH "Ground Truth Point", determine how well the "Candidate Pitch" covers it. You are grading on **Sales Effectiveness** and **Factual Precision**.

SCORING RUBRIC (0-5 Scale):

- **0 (Miss / Irrelevant):** The candidate pitch completely misses this concept. No mention of this feature, benefit, or metric.
- **1 (Marketing Fluff):** Vaguely mentions the topic (e.g., "improved efficiency") but lacks ANY specific substance. Critique: "This is a generic platitude that could apply to any company."
- **2 (Topic Match):** Identifies the correct Product or Pain Point, but misses the specific Solution or Outcome. *Example: GT says "Reduced downtime by 40%", Candidate says "Helps with downtime."*
- **3 (Implied / Soft Match):** Captures core value proposition correctly, but misses "Hero Evidence" (specific numbers, names, or unique mechanisms). *Verdict: A solid conversational point, but less persuasive than the Ground Truth.*
- **4 (Strong Sales Argument):** Captures the core value AND the key mechanism/outcome. It is a persuasive, accurate representation of the deal. *Difference from 5: Might miss a minor detail (e.g., date, exact city) that doesn't impact sales persuasion.*
- **5 (Strategic Bullseye):** A perfect extraction. Captures the **Product + Pain + Value + Specific Metric/Evidence** essentially verbatim from the Ground Truth. *Verdict: "This is exactly why they bought."*

OUTPUT FORMAT (JSON Only):

```
{
  "evaluations": [
    {
      "gt_point_id": <int>,
      "gt_summary": "<short_summary_of_gt_point>",
      "best_match_candidate_text": "<text_from_candidate_or_null>",
      "score": <0-5>,
      "reasoning": "<concise_sales_analysis>"
    }
  ]
}
```

Pitch Point Generation Prompt

I am a BDR at {seller} ({seller_website if seller_website else ""}). I want to sell {seller} Products to {customer} ({customer_website if customer_website else ""}). How should I pitch {"", ".join(products)} to {customer}? I need to generate 3 targeted pitch points and value propositions that address the customer's pain point which I can send to {customer} or their CEO/Leads/Decision Makers. Make sure to give me reasoning to each pitch point.

Requirements for each pitch point:

- Must be specific to the target company (use real facts you discover)
- Should connect the sender's products/services to the target's needs/pain points
- Must be atomic (one clear value proposition per point)
- Should be compelling and actionable

After researching, generate 3 pitch points which have a single value proposition per point addressing the customer's pain point. Each pitch point should be unique and not repeat the same value proposition. (it can repeat pitch points for the same product or/and pain point but the value proposition should always be different).

Respond with a JSON array of pitch point strings:

```
[
  "First pitch point text here",
  "Second pitch point text here",
  "Third pitch point text here",
  ...
  "Tenth pitch point text here"
]
```

Strictly stick to the above JSON format and structure.

Pitch Point Extraction Prompt

You are an expert sales analyst. Analyze the following success story content for {company_name} (the seller) and their product {product_name}.

Story Content:
{story_html}

Your goal is to extract specific, causal pitch points that explain WHY the product was successful, and provide EXACT QUOTES from the text as evidence.

STRICT FORMAT REQUIREMENT:

Return a JSON object with a single key "pitch_points" containing a list of objects. Each object must have the following structure:

```
{
  "summary": "<Product/Service> solved <Specific Pain Point> of <Customer> because
    of <Specific Value Proposition/Mechanism>[, resulting in <Quantitative Result>]",
  "evidence": [
    "Exact quote from the text supporting the pain point...",
    "Another exact quote..."
  ]
}
```

GUIDELINES:

1. **Product/Service:** Use the specific product name if available.
2. **Specific Pain Point:** What exact problem was the customer facing? Dig deeper than "inefficiency". Differentiate between "slow speed/manual work" and "inaccuracy/errors". These are DISTINCT pain points.
3. **Specific Value Proposition/Mechanism:** What specific feature or capability solved it?
4. **Quantitative Result:** IF the text mentions a number(e.g., "70% faster", "saved 10 hours"), YOU MUST INCLUDE IT in the summary.
5. **Evidence:** You MUST quote the original text. Do not paraphrase in this field.

Example Output:

```
{
  "pitch_points": [
    {
      "summary": "The Aldec G3 solved the high energy costs of T.Z. Osborne
        because of its PowerTubes technology that recovers kinetic energy,
        reducing consumption by 20%.",
      "evidence": ["Facing rising energy costs...", "The PowerTubes technology
        recovers kinetic energy, reducing consumption by 20%"]
    },
    {
      "summary": "Oracle Integration solved the issue of manual entry errors at
        Careem because of its automated GRN matching feature.",
      "evidence": ["Manual entry led to mismatches and compliance gaps.",
        "Automating GRN matching has significantly reduced errors."]
    }
  ]
}
```



A.8 Human Study to validate pitch point extraction

To validate the LLM’s accuracy and exhaustiveness in extracting pitch points, we conducted a human study on a random sample of 30 customer success stories. Annotators evaluated each story against the LLM-extracted pitch points along two dimensions: (1) **Precision** — verifying factual consistency and flagging hallucinations, and (2) **Recall** — identifying any pitch points the LLM missed. The LLM achieved a precision of **0.92**, recall of **0.97**, and an F1-score of **0.95**, validating its use as a robust, scalable proxy for ground truth extraction.

A.9 Institutional Review Board Approval

The human evaluation studies including the per-pitch usefulness field deployment with 12 sales development representatives and the gold-standard alignment exercise with senior SDRs from five enterprises were reviewed and approved by the Institutional Review Board. All participants were informed of the study’s purpose and provided consent prior to participation. No sensitive personal data was collected beyond professional judgments on model-generated sales content, and all responses were anonymized prior to analysis.

A.10 Statistical Significance and Confidence Intervals

We acknowledge that the evaluation is conducted on a limited subset due to the high computational cost of deep research agents. To ensure robustness, we perform bootstrapping (1,000 iterations) to compute 95% confidence intervals (CIs) for the Weighted Coverage Score (WCS) across the SDR-Bench evaluation set.

Table 6: Bootstrap Estimates of WCS on SDR-Bench Evaluation Set

Model	Mean WCS	95% CI
STORM	0.4246	[0.4015, 0.4491]
ODR	0.3358	[0.3150, 0.3585]
GPT-4o	0.3638	[0.3422, 0.3860]
Qwen	0.3692	[0.3496, 0.3876]

Validation of the Personalization Plateau. The 95% CIs for GPT-4o [0.3422, 0.3860] and Qwen [0.3496, 0.3876] exhibit substantial overlap, indicating no statistically significant difference in performance. This supports the existence of a *personalization plateau*, where different model architectures converge to a similar performance ceiling under our evaluation framework.

Significance of STORM. In contrast, the CI for STORM [0.4015, 0.4491] does not overlap with those of other models, indicating a statistically significant performance improvement.

Stability of Estimates. The relatively narrow width of the confidence intervals suggests stable estimates despite the limited sample size. The evaluation set comprises approximately 720 agent–environment interactions, providing a sufficiently representative estimate of model performance under the SDR-Arena setup.

A.11 Broader Impacts and Ethical Considerations

Our work raises important societal and ethical considerations, which we address below.

Paradox of Measurement and Dual-Use Risks. Benchmarking personalization creates an inherent tension: quantifying what makes sales outreach effective risks providing a blueprint for scalable, manipulative content. We argue, however, that the absence of transparent evaluation standards poses a greater risk by allowing opaque commercial systems to operate unchecked. **SDR-Arena** provides

the transparency needed to distinguish context-aware assistance from hallucinatory or manipulative outreach.

Privacy and Data Stewardship. Our dataset curation followed strict ethical guidelines. The proprietary email dataset was processed in a secure, access-controlled environment with all PII anonymized or redacted, and is *not* included in our public release. The public **SDR-Bench** is limited to already-published customer success stories, further filtered to enterprise entities to minimize individual exposure.

Economic Displacement and Human-AI Collaboration. Our findings reveal a “personalization plateau,” suggesting LLMs currently lag behind human experts in identifying nuanced, strategic revenue drivers. This supports a *Human-in-the-Loop* paradigm: our benchmark should guide assistants that reduce research drudgery for humans, not autonomous systems that replace human judgment.

Acceptable Use Policy. To mitigate the risks of misuse, the release of our framework and the **SDR-Bench** dataset will be accompanied by a restrictive Acceptable Use Policy. This policy explicitly prohibits the use of our artifacts or fine-tuned models for:

1. **Unsolicited High-Volume Outreach:** Using the dataset to train agents for mass-spamming or harassment.
2. **Deceptive Practices:** Generating content that masquerades as human correspondence without disclosure.
3. **Social Engineering:** Leveraging the personalization metrics to craft targeted phishing attacks.

By bringing scientific rigor to sales agent evaluation, we aim to steer the field toward personalization that respects user context and delivers genuine value, rather than optimizing for engagement at the expense of user trust.