Observed Appetite: A Computational Framework for Measuring Commercial Insurance Carrier Underwriting Behavior at Distribution Scale

Ankur Shrestha

QuoteSweep, United States

May 2026

Abstract

Determining which commercial insurance carriers will quote a given submission is a prerequisite to every commercial-lines placement, yet the signals available to independent agents — carrier-self-reported appetite guides, application-programming-interface (API) responses from a small set of integrated carriers, and individual agent experience — have not been formalized as a computational object, nor measured against each other at scale. We call this measurement task *observed appetite* and distinguish it from the three established categories: *stated appetite* (carrier-published), *API-revealed appetite* (returned by integrated platforms), and *verified appetite* (confirmed directly with the carrier). Observed appetite is the empirical distribution of a carrier's quote-or-decline behavior computed from real submission outcomes through that carrier's own portal, partitioned by industry classification, geography, and risk size.

We formalize observed appetite as a five-stage computational task (Submit → Classify → Aggregate → Score Confidence → Reconcile) with an explicit four-way ontology of appetite sources. We describe the QuoteSweep system, which implements the observation infrastructure through AI-powered browser agents operating on authenticated carrier portals without requiring API partnerships. We evaluate the *stated appetite* baseline against itself across a corpus of 509 commercial property and casualty (P&C) carriers. This paper presents the framework and the stated-appetite baseline; production-scale observed-appetite measurement is left to future work (§6.3). Only **2.2%** of carriers publicly disclose any industry × state × size interaction, only **1.2%** annotate appetite at six-digit NAICS resolution, and only **4.7%** of line-of-business commitments specify a revenue threshold. When the same carrier's published PDF appetite guide and its own public appetite web page are compared at the NAICS-2 sector level ($n = 189$ carriers, 3,780 cells), **the PDF asserts 2.14× more sector inclusions than the web page** (PDF "in" marginal 33.7% vs. 15.7%; Cohen's $\kappa = +0.25$, 95% CI 0.22–0.28 — fair agreement under Landis & Koch). Two surfaces of the same carrier describe systematically different coverage scopes. We release the corpus as an open dataset for reproducibility.

Keywords: observed appetite, commercial insurance, carrier appetite, browser agents, inter-rater agreement, insurance distribution, InsurTech, open data

1 Introduction

1.1 The market problem

Independent insurance agencies write approximately **87.2% of commercial-lines premiums** in the United States [1]. The commercial-lines market is large but not directly published as a single line item; we estimate it at roughly **$430 billion** annually, derived from the Insurance Information Institute's 2024 P&C net-written-premium total of $918.6B [2] and the IIABA-reported commercial

share. The carrier base across which this premium is distributed is the long tail of approximately **2,500 P&C insurers** in the United States (NAIC active-companies count, 2024; widely cited industry estimate) [3]. Every commercial submission an agent originates must be routed to one or more carriers willing to underwrite the risk. The agent's first computational task — and, in most agencies, an entirely manual one — is to determine *which* carriers fit *this* business.

The signals available to an agent are sparse and inconsistent. Carrier-published appetite guides, the historical primary source, are released on annual or quarterly schedules and describe appetite in coarse free-text language. Comparative-rating platforms with carrier API integrations [4, 5, 6] return structured availability and premium information but are limited to the small fraction of carriers willing to build, expose, and maintain an API. Industry-wide market-share data and direct conversations with underwriters round out the picture, but only for the carriers an individual agent has personally placed with before. The result is well documented. In First Connect Insurance's 2025 State of the Industry Report, **71%** of surveyed agents reported difficulty understanding carrier appetite, and **64%** reported quote-decline rates between 10% and 50% [7]. The most recent J.D. Power U.S. Independent Agent Satisfaction Study found that only **60%** of agents say their carriers clearly communicate risk appetite [8]. Industry analyses report that commercial underwriting teams thoroughly review only **30–40%** of incoming submissions, and that roughly 25% of submissions fall outside carriers' stated risk appetites [9].

1.2 Limitations of existing appetite signals

Three distinct technical regimes have produced industry tooling around this problem, none of which has been formalized as a measurement category.

Stated appetite is the carrier-self-reported category. Data products such as AskKodiak [10] and Ivans Markets aggregate appetite information that carriers themselves publish. The data structure is rich — AskKodiak's documentation describes mapping products against "over 20,000 data points" at six-digit NAICS resolution — but the underlying signal is whatever each carrier chooses to disclose. As we show in §5, what carriers actually disclose is dramatically coarser than the schemas built to receive it.

API-revealed appetite is the per-submission category. Platforms with carrier API integrations [4, 5, 6] return ground-truth quote-or-decline responses. The signal is high-fidelity within the integrated set, but the integrated set is small: the leading platform reached approximately 35 carrier partnerships after eight years of operation, a coverage rate of roughly 1.4% of the active P&C market [4].

Verified appetite is the bespoke category — appetite confirmed through direct underwriter conversation or through observation of consistent behavior over time. It is the most accurate signal available to a given agent and the most expensive to produce.

What has been missing is a fourth category: appetite measured at scale from observed quote-or-decline outcomes, across the carriers that lack API connectivity but maintain web portals — a category that, depending on the carrier set, can cover the majority of the market.

1.3 The technological enabler

The proximate enabler of this fourth category is the recent emergence of commercially viable AI-powered browser-agent infrastructure. Beginning with WebGPT [11] and the ReAct reasoning-and-acting pattern [12], a sequence of academic and commercial systems have demonstrated that language models, equipped with browser-control primitives, can navigate authenticated web interfaces and recover from failure with sufficient reliability to be useful in production. WebArena [13] and

Mind2Web [14] established reproducible benchmarks; WebVoyager [15], WorkArena [16], and Gur et al.'s real-world WebAgent [17] established that the same techniques transfer to realistic and authenticated environments. The performance gap to humans remains substantial — best published agents achieve 11.7–59.1% task success against human references near 90% [13, 15] — but the gap is small enough, on the *narrow* tasks that constitute commercial insurance quoting, to make portal-mediated quote acquisition operationally feasible.

This paper does not claim that browser agents have closed the human gap. It claims that the gap is now narrow enough that *observation* — recording the structured outcome of every browser-agent-mediated submission — produces a dataset whose marginal cost per row is low enough to support a continuously updated appetite estimator. The system described in §3 has been deployed against 509 distinct carriers in QuoteSweep's production environment; the cross-portal scrape that constitutes our stated-appetite baseline (§5) is itself a byproduct of that deployment.

1.4 Contributions

This paper makes four contributions, plus a narrower methodological contribution.

First, we coin **observed appetite** and distinguish it formally from stated, API-revealed, and verified appetite. The four-way ontology is presented in §2.1.

Second, we decompose observed appetite into a five-stage sequential computational task — Submit → Classify → Aggregate → Score Confidence → Reconcile — with explicit definitions of input, output, and reasoning challenge per stage (§2.3).

Third, we describe the QuoteSweep system that implements the framework. The system is the subject of two US Provisional Patent Applications: *System and Method for Automated Multi-Carrier Commercial Insurance Quoting Using Parallel AI Browser Agents Operating on Authenticated Insurance Carrier Web Portals Without Requiring Application Programming Interface (API) Partnerships* (Application 1) and *System and Method for Generating Predictive Insurance Carrier Appetite Intelligence by Observing and Analyzing Real-Time Quote Submission Outcomes from Automated Browser Agent Interactions with Insurance Carrier Web Portals* (Application 2), both filed by the author (§3).

Fourth, we evaluate stated appetite as a routing signal against itself. Across a corpus of 509 commercial P&C carriers, only 2.2% disclose industry × state × size interactions and only 1.2% annotate at six-digit NAICS resolution; same-carrier PDF guides assert 2.14× more sector inclusions than the carrier's own appetite web page (Cohen's $\kappa = +0.25$, fair agreement). We release the corpus under CC-BY-4.0 as supplementary material (§5, §6).

A fifth, narrower contribution is methodological: we report a *pipeline self-audit* (§3.6) that quantifies how much per-sector appetite signal QuoteSweep's normalization layer preserves from raw source text. Within insurance this is a footnote; outside it, the audit pattern — measuring agreement between a structured-extraction pipeline and the human-readable source it was extracted from — generalizes to any LLM-mediated extraction system.

2 Formalizing Observed Appetite

2.1 Instrument ontology: four categories of appetite data

Underwriting appetite is not a single object. We distinguish four categories by the *act* that produces the data, as summarized in Table 1.

The four categories are *not* mutually exclusive on a given carrier-NAICS-state cell — a carrier may simultaneously have a stated-appetite record, an API-revealed quote rate, an observed quote

Table 1: Four categories of commercial-insurance carrier appetite data, distinguished by producing act. Observed appetite is the contribution of this paper.

Category	Producing act	Cadence	Coverage	Bias profile
Stated appetite	Carrier self-publishes guide / API record	Quarterly to annual	≈ all 2,500 P&C carriers in principle; uneven in practice	Marketing-direction bias; temporal lag; granularity loss (this paper, §5)
API-revealed appetite	Carrier API returns quote/decline	Per submission	≈ 35 carriers via leading platform [4]; ≈ 1.4% of market	Honest within scope; bounded by partnership count
Observed appetite *(this paper)*	Carrier portal returns quote/decline to a browser-agent-mediated submission	Continuous	Any carrier with a web portal (≈ high-90% of market)*	Cold-start sparsity; portal-side selection effects
Verified appetite	Direct underwriter confirmation or high-volume observed agreement	As needed	Sparse	Authoritative; expensive

*Derived as the complement of the small fraction of carriers covered by API rater integrations; the actual rate is bounded above by the share of active P&C carriers that maintain any web portal.

rate, and a verified appetite communication. They differ in producing act and in the inferential structure each supports.

2.2 Observed appetite as a distinct computational task

We define **observed appetite** as the task of inferring, for each (carrier c, industry classification n, state s) tuple, the empirical distribution

$$\hat{q}_{c,n,s} \;=\; \mathbb{P}\big(\,\text{quote} \mid c, n, s, \text{submission complete}\,\big) \tag{1}$$

from accumulated observation records, where "submission complete" denotes outcomes in {quoted, declined-appetite, declined-other} — i.e., the carrier rendered an underwriting decision. Technical failures and timeouts are excluded from both numerator and denominator. This is conceptually a domain-specific instance of revealed-preference inference [18, 19, 20] — preferences (here, underwriting policy) are read from actions (quote, decline) rather than from declarations (published guides). The motivating intuition is closely related to the asymmetric-information tradition in insurance economics [21, 22]: where the carrier-side underwriting policy is partially observable and partially private, observation of actual quote-or-decline behavior is a more reliable preference signal than self-published guides. Two properties distinguish observed appetite from adjacent computational tasks.

First, **observed appetite requires action across organizational boundaries.** Within-carrier behavioral modeling — predicting which submissions a carrier's own underwriters will

bind, given their historical book — is a well-developed line of work, including the analytics agenda described in McKinsey's "From art to science" survey [23] and the personal-lines telematics literature. Cross-carrier appetite measurement, by contrast, requires *probing* the carrier from the outside, because no individual carrier has the cross-carrier book that would let it answer the question internally.

Second, **observed appetite has a cold-start problem with a curated fallback.** New carriers, new NAICS codes, and new (carrier, NAICS, state) cells begin with zero observations. The cold-start literature [24] provides primitives for sparse-cell estimation; here, the natural fallback is the carrier's stated appetite — exactly the signal whose limitations motivated this paper. We formalize the trade-off in §2.3 stage 4 (Score Confidence) and analyze its operating regime empirically in §5.

2.3 The five-stage decomposition

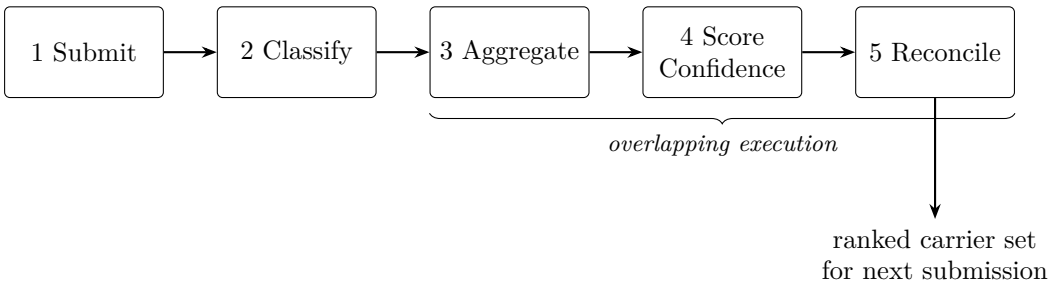

Figure 1: The five-stage observed-appetite pipeline. Each stage is logically sequential; in practice stages 1–2 execute per submission and stages 3–5 update continuously as observations accumulate.

Stage 1 — Submit. Input: a normalized business risk record (NAICS code, state, annual revenue, employee count, years in business, requested line of business) and the agent's credentials for each target carrier portal. Output: a terminal portal state per (carrier, line-of-business) submission — a results page, a decline page, a technical error, or a timeout. The computational challenge is navigating heterogeneous authenticated portals without selector-brittle DOM scripting; recovering from MFA, conditional forms, and timeouts; and reasoning about portal state from a natural-language prompt. Web-agent benchmarks [14, 13, 25, 16, 15, 17] benchmark this capability on consumer tasks; commercial insurance portals are authenticated B2B financial workflows with multi-class terminal states and asymmetric failure costs.

Stage 2 — Classify. Input: the terminal portal artefact (HTML and screenshot). Output: a typed observation record with outcome $\in$ {quoted, declined-appetite, declined-other, technical-failure, timeout}; premium (if quoted); decline reason text (if declined); and the anonymized risk characteristics. The challenge is distinguishing *appetite declinations* (underwriting refused the risk) from *technical failures* (the session never reached an underwriting decision). Misclassifying technical failure as appetite declination corrupts the appetite estimator downstream. Application 2 [26] describes a 15-keyword first-generation detector; a learned multi-class classifier is the natural extension as observation volume grows.

Stage 3 — Aggregate. Input: the stream of typed observation records over time. Output: a time-windowed empirical quote-rate estimate $\hat{q}_{c,n,s}$ per cell, plus a trend classification {improving, stable, declining, insufficient_data}. The challenge is sparse-cell inference, sufficient-statistics thresholds, and trend detection across sequential windows. Application 2 sets the minimum threshold at five non-technical observations per cell and the trend window at three months. Recommender-

system cold-start [24] provides methodological primitives; cross-carrier insurance applications are, to our knowledge, unstudied as a class.

Stage 4 — Score Confidence. Input: an appetite estimate for a cell, with provenance — manual rule, observed with N observations, or verified. Output: a confidence-tier annotation and a precedence-aware composite estimate when sources conflict. Application 2 specifies a strict ordering: manual < observed < verified, with observed treated as supplementary when $N < 10$ and as a candidate for automated rule promotion when $N \geq 50$. The deeper Bayesian question — when does N observations justify overriding a manual rule from a published underwriting guide? — is addressed in current implementations by a fixed threshold and is the natural locus for future learned weighting.

Stage 5 — Reconcile. Input: the composite estimate from Stage 4 plus a new submission. Output: a ranked carrier list with annotations: in-appetite; in-appetite-with-low-historical-quote-rate-warning; out-of-static-appetite-but-historically-quoted; out-of-appetite. Application 2 describes the asymmetric override semantics — *upgrade* a static decline if observation is strong enough ($\geq 80\%$ observed quote rate), *retain-but-flag* a static accept if observation is weak ($\leq 20\%$ observed quote rate). The asymmetric thresholds are a specific instance of a more general problem: how to communicate uncertainty to a user while preserving actionability.

2.4 Why observation-based measurement requires browser-agent infrastructure

A reviewer might reasonably ask why observed appetite could not be produced by aggregating quote outcomes from existing API-integrated platforms. The answer is the API coverage ceiling: across all major commercial-lines API rater platforms surveyed in §4.3, the union of integrated carriers is bounded at a small fraction of the active P&C market. To measure observed appetite at the scale where the long tail of carriers — small mutuals, regional specialists, surplus-lines writers — becomes statistically observable, the data-acquisition mechanism must not depend on bilateral carrier cooperation. Browser-agent infrastructure makes that data-acquisition mechanism economically and operationally feasible for the first time.

3 System Design

We describe the QuoteSweep architecture that implements the five-stage pipeline. The system is the subject of two US Provisional Patent Applications filed by the author, hereafter cited as Application 1 [27] (multi-carrier quoting infrastructure) and Application 2 [26] (appetite intelligence layer). Implementation details below summarize the technical disclosure in those filings.

3.1 Goal-prompt navigation (Stage 1)

For each (carrier, line-of-business) pair the system supports, the registry stores a structured natural-language **goal prompt** with four sections: TASK (the high-level objective), FILL (form field mappings, populated by template interpolation from the submission record), GUARDRAILS (handlers for popups, cookie banners, navigation errors, timeouts), and EXTRACTION (a multi-strategy pipeline for premium extraction: direct field $\rightarrow$ monthly $\times$ 12 annualization $\rightarrow$ regex fallback). The prompt is interpreted by an AI browser agent — currently routed through a configurable cascade of commercial providers — that adapts to the current state of the portal without depending on CSS selectors or XPath expressions. The structure is, in effect, a domain-specific instance of the ReAct reasoning-and-acting pattern [12].

The system implements a cascading multi-provider failover (Application 1 claim 3): a primary provider handles each job; on failure the job cascades to a secondary, then a tertiary, with provider selection controlled by environment-variable configuration. This decouples the system from any single browser-automation vendor.

3.2 Observation schema (Stage 2)

Every completed Stage-1 submission emits an observation record with the schema specified in Application 2: `{carrierId, lineOfBusiness, naicsCode, state, annualRevenue, yearEstablished, numberOfEmployees, outcome, premiumReturned, declineReason, quoteJobId, agencyId, observedAt}`.

The schema deliberately excludes business names, addresses, contact information, and all other personally identifiable business information; this constraint is structural, not configurable, and is intended to enable aggregation across agencies for cross-carrier intelligence without exposing individual insureds.

3.3 Declination detection (Stage 2)

A 15-keyword classifier separates appetite declinations from technical failures (Application 2 claim 2). Appetite keywords include "unable to provide a quote," "does not meet our underwriting guidelines," "not eligible for online quoting," "outside our appetite," and several variants. Technical-failure indicators include session-timeout codes, CAPTCHA challenges, and portal-maintenance pages. The classifier is intentionally simple at this stage; the observation table provides the labeled data on which a learned successor can be trained.

3.4 Confidence hierarchy and blended evaluation (Stages 4–5)

The system implements the hierarchical confidence classification described in Application 2: manual $<$ observed $<$ verified. The blended-evaluation engine combines static appetite rules with the observed quote rate using the asymmetric thresholds specified in §2.3 stage 5. Override events — cases in which an agent submits a quote to a carrier flagged out-of-appetite — are themselves logged as a metric on the appetite system, providing a downstream signal that the rule logic is over- or under-restricting.

3.5 Privacy-preserving design

Per Application 2 claim 1(c), no element of the observation record is personally identifiable to a specific business or insured. Carrier portal credentials — including usernames, passwords, and TOTP shared secrets — are vaulted in AWS Secrets Manager under envelope encryption by a customer-managed AWS KMS key (AES-256-GCM), with the application database retaining only an opaque ARN reference. Plaintext credential material never resides in the operational database and is retrieved on demand through IAM-authenticated SDK calls scoped per environment; time-based one-time passwords are generated server-side from the vaulted shared secret per RFC 6238. Inter-agency data isolation is enforced at the row level on the credential and submission tables. The intent is twofold: to support eventual third-party data products (carrier-side appetite analytics, market-trend studies) without privacy exposure on the agent or insured side, and to make the observation dataset releasable in aggregate form for academic use.

3.6 Pipeline self-audit

A system whose output is structured data has a known but rarely measured failure mode: information loss in the extraction-and-normalization layer between the raw source artefact and the structured record. We report a self-audit of QuoteSweep's normalization layer in Table 2. Twenty-five carriers were sampled stratified across `sourceType`; for each, a 20-NAICS-2-sector inclusion vector was coded independently from the raw source text and from the carrier's parsed JSON record (the input to Stage 3). Pooled agreement is summarized below.

Table 2: Pipeline self-audit. AC1 is reported as the appropriate prevalence-robust statistic; standard Cohen's κ is paradox-affected because both raters mark "in" the majority of cells. The 77.7% preservation rate is the share of source-asserted sector inclusions that the parse retains; the 22.3% loss rate is the share dropped or altered in normalization.

Quantity	**Value**
Cells coded	161
Raw observed agreement	67.1%
Cohen's κ	-0.190 (paradox-affected; see [28] for discussion)
Gwet's AC1 [28]	$+0.546$ [95% CI $+0.404$, $+0.663$]
Pipeline preservation rate	77.7%
Pipeline loss rate	22.3%

We surface this number explicitly because pipeline self-audit is methodologically valuable in any LLM-mediated structured-extraction system and is rarely reported. The corollary is that any inference downstream of Stage 3 inherits the 22.3% loss as an upper bound on signal fidelity — a hedge that the existence of a measured baseline makes possible.

4 Related Work

We position observed appetite within four adjacent literatures, three academic and one industrial.

4.1 Web and browser agents

The technical substrate of Stage 1 is the contemporary line of work on LLM-powered web agents. ReAct [12] introduced the reasoning-and-acting interleaving pattern that goal-prompt navigation generalizes. Toolformer [29] established self-supervised tool-use; WebGPT [11] demonstrated browser-as-tool at scale. Mind2Web [14] introduced the first large open dataset of real-world web tasks (2,000+ tasks across 137 websites); WebArena [13] and VisualWebArena [25] established the benchmarks against which current agents are measured (best published task-success rates of 11.7% and 16.4% respectively, against human references of 78.2% and 88.7%). WorkArena [16] and BrowserGym extend benchmarking to authenticated enterprise workflows. WebVoyager [15] reports 59.1% task success on real websites using a multimodal architecture similar in spirit to the goal-prompt approach. Gur et al.'s WebAgent [17] reports 70% success on real-world navigation with planning, long-context HTML summarization, and program synthesis. AgentBench [30] provides a multi-environment evaluation framework.

The observed-appetite literature inherits the techniques but addresses a different distributional problem: authenticated B2B financial workflows with multi-class terminal states and asymmetric failure costs. The published web-agent benchmarks evaluate consumer tasks (shopping, booking)

where success is binary and failure is recoverable; commercial-insurance quoting is neither. We make no claim that observed-appetite measurement requires web agents to reach human-level reliability on those benchmarks; rather, the operational claim is that current published systems are reliable enough on narrow, scripted commercial-portal workflows to produce usable observations at scale.

4.2 InsurTech research and the appetite gap

The academic InsurTech literature is concentrated on (a) risk pricing and classification, (b) claims and fraud prediction, (c) personal-lines telematics, and (d) digital-platform business models. Appetite as a *measurement object across carriers* is essentially absent. The closest adjacent work is the within-carrier production-side analytics agenda surveyed by McKinsey ("From art to science" [23]) — but that agenda treats appetite as a carrier's internal policy to be operationalized via its own data, not as an externally observable phenomenon to be measured across carriers. The 2025 InsurTech special issue of the *Geneva Papers on Risk and Insurance* surveys digital technology in insurance broadly without addressing appetite measurement specifically.

We interpret this gap as the white space the present paper occupies: the cross-carrier appetite-as-measurement-object question has been a practitioner concern [8, 7] without a corresponding academic formalization.

4.3 Existing appetite data products

Three commercial categories of appetite tooling exist in production, none of which has published systematic evaluation results.

The *stated appetite* category is dominated by AskKodiak (now part of Ivans / Applied Systems) [10], which provides a carrier-facing self-publication tool that maps products against a six-digit NAICS schema with associated underwriting metadata. Carriers populate their own records; Ivans' product page describes the data as "real-time" but the actual update cadence depends on each carrier's editorial discipline. Our §5 analysis is consistent with the interpretation that the published-versus-promised gap is substantial.

The *API-revealed appetite* category includes Tarmika (Applied Systems) [4], Semsee, Bold Penguin [5], and Vertafore Commercial Submissions [6]. Tarmika, the most-cited example, reports approximately 31–35 carrier integrations after eight years of operation; Bold Penguin reports 40+ integrated carriers on its carrier platform plus 45+ specialty markets via the SquareRisk acquisition; Vertafore reports 25+ carriers across five commercial lines. None publishes carrier-by-carrier agreement metrics. The aggregate observable result — adoption rates of 24% (any commercial rater) and 26% (Tarmika specifically) among surveyed independent agencies [31] — is consistent with utility within the integrated set being high while the integrated set itself remains small.

Agentero recently announced (November 2025) an "AI Appetite Checker" that combines carrier-published appetite guides with the platform's own bound-policy data, scoped to Agentero's agent network [32]. This is the closest commercial framing to observed appetite to date and is, in our reading, evidence that the category is forming. The distinction relevant to this paper: Agentero's signal is bounded by the carriers and agents on Agentero's network; observed appetite, as defined in §2, is bounded by the substantially larger set of carriers that maintain a web portal.

4.4 The industrial-academic gap

The pattern in §4.2–§4.3 — active industrial systems addressing aspects of a problem that has not been academically formalized — is not unique to insurance. Halioua, Bloch, and Guez [33] make the same observation about regulatory compliance in their MARIA framework, which formalizes

"regulatory intelligence" as a computational task distinct from compliance checking, obligation extraction, and regulatory change detection. The structural parallels are worth noting because they are common to industrial categories that emerge faster than the corresponding academic literature: a measurement object is implicit in many production systems, but the explicit definition, evaluation protocol, and error taxonomy are produced late or not at all.

This paper adopts the MARIA-style move for commercial-insurance distribution: define the object, decompose the task, evaluate the existing baseline against itself, and release a corpus that supports reproducible follow-on work.

5 Empirical Analysis of Stated Appetite

We evaluate the stated-appetite baseline against itself. The motivating question is whether stated appetite, as actually published by carriers, is internally consistent and granular enough to function as a submission-routing signal independent of observation.

5.1 Dataset and methodology

The corpus consists of 509 commercial P&C carriers whose appetite documentation was collected through QuoteSweep's discovery pipeline between March and April 2026 [34]. For each carrier, the corpus contains a primary source artefact — a carrier-published PDF guide ($n = 201$) or a text-page scrape of the carrier's public appetite or industries page ($n = 308$) — plus a normalized JSON record produced by an LLM-assisted parser. Both layers are released under CC-BY-4.0 in supplementary material at the paper's permanent URL.

Two complementary studies are reported. **Analysis A** measures inter-source agreement on which industry sectors a carrier writes, computed between two independent public surfaces of the *same* carrier. **Analysis B** measures the structural granularity of stated appetite at the carrier level along six dimensions (industry granularity, state granularity, size threshold, exclusions, interaction disclosure, date disclosure). The full coding protocol, verbatim evidence quotes for the interaction-disclosure dimension, reproducible Python scripts (seed `20260518`), and bootstrap details (1,000 resamples for all confidence intervals) are in supplementary material.

5.2 Same-carrier coverage asymmetry across public surfaces (Analysis A)

5.2.1 Methodology

Of the 509 carriers in the corpus, 189 carriers have **both** a carrier-published PDF appetite guide *and* a separate appetite-page text scrape on disk — that is, two distinct public surfaces on which the same carrier expresses its appetite. For each such carrier, the unit of analysis is the 20-NAICS-2-sector inclusion vector. The two raters being compared are the same physical text extracted from two surfaces (PDF, text page); the same deterministic sector-keyword matcher is applied to each. The category set is binary: a sector is coded `in` if the source text mentions the carrier's appetite for that sector (matched against a fixed list of NAICS-2 sector keywords), and `out` otherwise. **All 20 sectors per carrier are retained**; we do not drop cells. The total sample is $189 \times 20 = 3{,}780$ cells.

Cohen's κ [35] is computed in the standard form

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

where p_o is the observed agreement rate and p_e is the chance agreement rate computed from the rater marginals. Confidence intervals are 1,000-resample bootstrap with seed `20260518`. Gwet's AC1 [28] is reported alongside κ as a prevalence-robust statistic; for binary data,

$$\text{AC1} = \frac{p_o - 2\bar{\pi}(1-\bar{\pi})}{1 - 2\bar{\pi}(1-\bar{\pi})}, \quad \bar{\pi} = \frac{p_{\text{PDF, in}} + p_{\text{TXT, in}}}{2}. \tag{3}$$

5.2.2 Confusion matrix and pooled result

Table 3: Contingency table for same-carrier PDF vs. appetite-web-page sector codings, pooled across 189 carriers × 20 NAICS-2 sectors.

	TXT "in"	TXT "out"	**PDF total**
PDF "in"	384	888	1,272
PDF "out"	210	2,298	2,508
TXT total	**594**	**3,186**	**3,780**

Pooled statistics:

Table 4: Pooled inter-source agreement and marginal statistics for the 189-carrier within-carrier comparison.

Quantity	**Value**
Carriers paired	189
Cells coded	3,780
Raw observed agreement p_o	0.7095 (70.95%)
Chance agreement p_e	0.6121
Cohen's κ	+0.2511 [95% bootstrap CI 0.2216, 0.2821]
Gwet's AC1 [28]	+0.5376 [95% CI 0.5101, 0.5652]
PDF source — "in" marginal	**33.65%**
Text-page source — "in" marginal	**15.71%**
PDF-to-TXT coverage ratio	**2.14**×
Landis & Koch interpretation [36]	*fair agreement* (0.21–0.40)

5.2.3 Interpretation: coverage asymmetry, not contradiction

The $\kappa = +0.25$ result indicates fair agreement; this is meaningful alignment but a long way from substantial or near-perfect. The more substantive observation, however, is in the marginals. **A carrier's published PDF guide asserts the carrier writes 2.14× more NAICS-2 sectors than the same carrier's appetite web page does.** Of the 1,098 cells where the two sources disagree, 80.9% are PDF-in / TXT-out (the PDF is broader) and 19.1% are PDF-out / TXT-in. The pattern is consistent with the two surfaces serving systematically different communication functions — the PDF appears in agent-resource libraries and tends to enumerate the carrier's full long-tail appetite; the web page appears to prospective insureds and tends to surface only the carrier's primary segments.

The implication for routing is not that stated appetite is internally contradictory, but that *which surface an agent reads* materially changes the carrier set the agent would consider. An agent relying on the carrier's web page would consider a substantially narrower carrier-sector set than one reading the same carrier's PDF guide. Neither surface is dispositive of the carrier's actual underwriting behavior; both are upstream of observation.

5.2.4 Sensitivity check: the cost of a non-standard drop rule

An initial computation (preserved in supplementary file `analysis-a-study1.json`) applied a drop rule that excluded cells where both sources coded "out," on the reasoning that such cells were "uninformative." This rule excluded 2,298/3,780 = 60.8% of cells — including precisely the both-agree-out cells that anchor κ's chance-correction baseline. Under that drop rule, the same data yields $\kappa = -0.30$ ("poor"). We report this as a sensitivity check, not as a finding: the drop rule is non-standard for 2×2 κ on a fixed sector grid, and the swing from $+0.25$ to -0.30 by adding or removing the both-out cells illustrates why the full grid must be retained. Both numbers are released in supplementary material; the $+0.25$ figure is the headline reported here.

5.3 The granularity gap (Analysis B)

We coded each of the 509 carriers along six structural dimensions. The headline metrics are in Table 5; full distributional results are in supplementary material.

Table 5: Granularity-gap metrics across the 509-carrier corpus. Confidence intervals are 1,000-resample bootstrap. All metrics computed by the scripts released as supplementary material.

Metric	**Result**	n / N
D5 Industry × state × size interaction disclosed	**2.16%** [95% CI 0.98, 3.54]	11 / 509
D3 Revenue threshold disclosed (LOB-row)	**4.68%** [95% CI 3.79, 5.61]	95 / 2,031
D1 NAICS-6 industry resolution	**1.18%** [95% CI 0.20, 2.16]	6 / 509
D6 Explicit date on guide	22.40% [95% CI 18.86, 25.93]	114 / 509
D4 Explicit exclusion list	30.85%	157 / 509
Full three-dimension disclosure (D1 ∧ D2 ∧ D3 ∧ D5)	**1.77%** [95% CI 0.79, 2.95]	9 / 509
LOB rows with `statesAvailable = "all"`	73.97%	1,290 / 1,744
`guideYear` field missing on parsed record	73.48%	374 / 509

The dominant industry-classification ontology in the corpus is *free-text English* (472 / 509 carriers = 92.7%); 6 / 509 use class codes (SIC or NCCI), 6 / 509 use NAICS at six-digit resolution, and 25 / 509 do not disclose industry information in machine-readable form. The discrepancy with the schema used by leading stated-appetite data products [10] — which is designed to ingest NAICS-6 — is the proximate explanation for the inter-source disagreement reported in §5.2: when the source data is overwhelmingly free-text and the target schema is NAICS-6, the normalization step is itself the dominant source of variance.

A 30-carrier manual audit of D5 (interaction disclosure) negatives identified 1 borderline false negative under a more lenient coding rule, raising the upper bound on D5 = yes to 12 / 509 (2.36%). The conclusion is robust to either coding.

5.4 Discussion

Both analyses converge on a single substantive conclusion: stated appetite, in the form currently published by US commercial P&C carriers, is too coarse — and, across a carrier's own public surfaces, too inconsistently scoped — to function as an independent submission-routing signal at carrier scale. The 2.14× coverage asymmetry between a carrier's PDF guide and its own appetite web page (§5.2) is itself a useful operational observation: agents shown different surfaces of the same carrier would consider materially different carrier sets. The 2.2% interaction-disclosure rate (§5.3) is, in our reading, the more fundamental limitation: even the best-disclosed surface generally cannot answer the underwriting-relevant question "in this state, for this industry, at this revenue band, would this carrier quote?"

This conclusion does not imply that stated appetite is *useless* — it is the only signal available for cells with zero observations, and it is the editorial substrate from which observed appetite is bootstrapped (§2.3 Stage 4). The argument is narrower: stated appetite alone is insufficient. The path to a usable cross-carrier appetite signal at scale requires either dramatically improved carrier-published disclosure — for which no current trend evidence exists in the corpus — or a second signal source. Observed appetite is one such source.

A self-aware limitation: the corpus is sampled from carriers that *publish* appetite documentation at all. The headline disclosure rates (2.2%, 1.2%, 4.7%) are therefore ceilings; the population mean across all 2,500 P&C carriers, weighted by the long tail of carriers that publish nothing, is almost certainly lower.

6 Discussion

6.1 Limitations

Six limitations:

Empirical scope. The empirical analyses in §5 use only stated-appetite public data. The paper does not report observed-appetite production figures, because the production observation set has not yet reached the volume and stratification at which per-cell quote rates would be reliable. The framework is in deployment; the production-scale evaluation is left to future work (§6.3).

Single-coder coding. All coding in §5 was performed by one LLM agent in one session. Every D5 = yes cell carries a verbatim evidence quote in the released CSV; a second-coder human validation pass is recommended before strong reliance on the per-cell codings. The aggregate metrics are robust to plausible coding noise; the within-cell judgments are not.

Sampling bias. The 509-carrier corpus over-represents carriers that publish *any* appetite documentation. Carriers that publish nothing are absent. The disclosure rates reported in §5.3 should therefore be read as upper bounds on the population, not as central estimates.

Pipeline preservation. The pipeline self-audit (§3.6) reports a 77.7% preservation rate from raw source text to normalized JSON. This is an upper bound on the fidelity of any downstream inference that depends on the parsed records — including parts of Analysis B. The reported metrics are still meaningful, but a reader interested in the absolute level of granularity disclosure should treat the parsed JSON as a slightly noisy proxy for the raw source.

Kappa paradox in the pipeline audit. Cohen's κ in §3.6 is paradox-affected because both raters mark cells "in" the majority of the time. Gwet's AC1 is reported as the appropriate prevalence-robust statistic [28]. The Analysis A Study 1 result in §5.2 does not exhibit the paradox condition — the two raters' marginals differ substantially (33.7% vs. 15.7% "in") — and the standard κ is reported as the headline there.

Methodology transparency on §5.2. An initial internal computation applied a non-standard drop rule that excluded both-out cells; under that rule the same data yields $\kappa = -0.30$, which would have been a misleading headline. The published analysis retains the full 20-sector grid per carrier, yielding $\kappa = +0.25$. Both numbers are released in supplementary material so the methodology choice is auditable; the +0.25 figure is the one we report.

6.2 Methodological findings that generalize beyond insurance

Two findings from this work generalize beyond commercial insurance distribution.

First, **same-source inter-surface measurement is a useful diagnostic for any organization that publishes the same information through multiple channels.** The 2.14× coverage asymmetry reported in §5.2 was not produced by comparing carriers to each other; it was produced by comparing each carrier's own public surfaces to each other. Any organization whose product, pricing, or policy information appears in multiple public locations can be measured by the same protocol. We expect the result generalizes to other industries with under-coordinated multi-channel publication — regulatory filings, financial product descriptions, healthcare-coverage policies.

Second, **pipeline self-audit (§3.6) is a credibility-cheap and information-rich addition to any LLM-mediated structured-extraction system.** Reporting a measured signal-preservation rate against the source text — rather than asserting that the parse is correct — bounds the trust a reader can place on downstream metrics and is, in our experience, a small fraction of the engineering effort of building the parser in the first place. The pattern generalizes to other LLM-assisted-extraction systems.

6.3 Scope and future directions

The system described in §3 is deployed in production within QuoteSweep. As of this filing, the observation table contains a non-trivial volume of submission-outcome records, but the per-cell density at the resolution required to publish observed quote rates with bootstrap confidence intervals is not yet sufficient across the long tail of (carrier × NAICS × state) cells. Future work, planned as a separate empirical paper, will report:

1. Production-scale empirical observed appetite for the cells with sufficient density.
2. The agreement between observed quote rates and the stated-appetite baseline reported here — and, where they diverge, the directional pattern of divergence.
3. Comparison of observed appetite to the API-revealed appetite of the small set of carriers covered by both QuoteSweep and an existing API rater.
4. Updated pipeline self-audit at larger stratified sample.

The path from the present framework to that follow-on empirical work is operational, not algorithmic; we do not claim it as a research contribution.

6.4 Open data and reproducibility

The 509-carrier corpus is archived on Zenodo under CC-BY-4.0 with DOI 10.5281/zenodo.20280436; a mirror is hosted at `https://quotesweep.com/research/observed-appetite/`. The release contains:

- The normalized JSON records used in §5.

- A per-carrier source map listing the original URLs and access dates (508 / 509 with URL; 459 / 509 with access date).
- A codebook documenting every field, with known gaps.
- The Python scripts that reproduce every metric in §5 (seed `20260518`).
- The verbatim D5 evidence quotes referenced in §5.3.

We do not release agent credentials, portal automation prompts, or any data that would identify specific submissions, agents, or insureds.

7 Conclusion

We have proposed *observed appetite* as a fourth category of commercial-insurance carrier appetite measurement, complementing the three categories — stated, API-revealed, and verified — that the industry currently uses. We have formalized it as a five-stage computational task with explicit per-stage inputs, outputs, and challenges; described the QuoteSweep architecture that implements the framework and is the subject of two pending patent applications; and evaluated the stated-appetite baseline against itself across a corpus of 509 carriers. The headline result — that only 2.2% of carriers publicly disclose any industry × state × size interaction, only 1.2% annotate appetite at six-digit NAICS resolution, and a carrier's own published PDF asserts 2.14× more sector breadth than its own appetite web page — establishes that stated appetite is insufficient as a routing signal and motivates the construction of a complementary observation-based measurement layer.

Follow-on empirical work with production-scale observed-appetite data is planned. The corpus underlying the empirical analyses here is released as open data.

Acknowledgments

The author thanks the early operators and independent insurance agents who provided feedback on production deployment of the system described in §3. Errors and interpretation are the author's alone.

References

[1] Independent Insurance Agents and Brokers of America (IIABA, Big "I"), "2025 Market Share Report," IIABA, Tech. Rep., Jul. 2025, coverage: *Insurance Journal*.

[2] Insurance Information Institute, "Facts + statistics: Industry overview; facts + statistics: Commercial lines," Available: https://www.iii.org/fact-statistic/facts-statistics-industry-overview, 2025.

[3] National Association of Insurance Commissioners (NAIC), "2024 Annual Property/Casualty and Title Insurance Industries Analysis Report," NAIC, Tech. Rep., Mar. 2025.

[4] Applied Systems, "Tarmika: The commercial quoting tool designed for your agency," Available: https://www.tarmika.com/, 2025.

[5] Bold Penguin, "Small business insurance companies, carriers, and insurance mga," Available: https://www.boldpenguin.com/, 2025.

[6] Vertafore, "Commercial submissions for insurance brokers and carriers," Available: https://www.vertafore.com/products/commercial-rater/commercial-submissions, 2025.

[7] First Connect Insurance, "2025 State of the Industry Report," Public blog summary: https://www.firstconnectinsurance.com/blog/2025-state-of-the-industry-report/. Full report (containing the 71% / 64% / 86% figures) is email-gated at https://info.firstconnectinsurance.com/state-of-the-industry-report-2025, 2025.

[8] J.D. Power, "2025 U.S. Independent Agent Satisfaction Study," Press release. Available: https://www.jdpower.com/business/press-releases/2025-us-independent-agent-satisfaction-study. Secondary summary: *IA Magazine*, Sep. 2025, 2025.

[9] SortSpoke, "Insurance submission triage explained," Available: https://sortspoke.com/blog/underwriting-submission-triage-explained, 2025.

[10] Ivans (Applied Systems), "Ask Kodiak for Carriers product page," Available: https://www.ivans.com/for-carriers/products/ask-kodiak/, 2025.

[11] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, "WebGPT: Browser-assisted question-answering with human feedback," 2021, arXiv:2112.09332.

[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *Proc. International Conference on Learning Representations (ICLR)*, 2023, arXiv:2210.03629.

[13] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig, "WebArena: A realistic web environment for building autonomous agents," in *Proc. International Conference on Learning Representations (ICLR)*, 2024, arXiv:2307.13854.

[14] X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su, "Mind2Web: Towards a generalist agent for the web," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2023, arXiv:2306.06070.

[15] H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, and D. Yu, "WebVoyager: Building an end-to-end web agent with large multimodal models," in *Proc. Association for Computational Linguistics (ACL)*, 2024, arXiv:2401.13919.

[16] A. Drouin, M. Gasse, M. Caccia, I. H. Laradji, M. Del Verme, T. Marty, L. Boisvert, M. Thakkar, Q. Cappart, D. Vázquez, N. Chapados, and A. Lacoste, "WorkArena: How capable are web agents at solving common knowledge work tasks?" in *Proc. International Conference on Machine Learning (ICML)*, 2024, arXiv:2403.07718.

[17] I. Gur, H. Furuta, A. Huang, M. Safdari, Y. Matsuo, D. Eck, and A. Faust, "A real-world WebAgent with planning, long context understanding, and program synthesis," 2023, arXiv:2307.12856.

[18] P. A. Samuelson, "A note on the pure theory of consumer's behaviour," *Economica*, vol. 5, no. 17, pp. 61–71, 1938.

[19] S. N. Afriat, "The construction of utility functions from expenditure data," *International Economic Review*, vol. 8, no. 1, pp. 67–77, 1967.

[20] H. R. Varian, "The nonparametric approach to demand analysis," *Econometrica*, vol. 50, no. 4, pp. 945–973, 1982.

[21] G. A. Akerlof, "The market for "lemons": Quality uncertainty and the market mechanism," *Quarterly Journal of Economics*, vol. 84, no. 3, pp. 488–500, 1970.

[22] M. Rothschild and J. Stiglitz, "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information," *Quarterly Journal of Economics*, vol. 90, no. 4, pp. 629–649, 1976.

[23] McKinsey & Company, "From art to science: The future of underwriting in commercial P&C insurance," McKinsey & Company, Tech. Rep., Sep. 2021.

[24] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. ACM SIGIR*, 2002, pp. 253–260.

[25] J. Y. Koh, R. Lo, L. Jang, V. Duvvur, M. C. Lim, P.-Y. Huang, G. Neubig, S. Zhou, R. Salakhutdinov, and D. Fried, "VisualWebArena: Evaluating multimodal agents on realistic visual web tasks," in *Proc. Association for Computational Linguistics (ACL)*, 2024, arXiv:2401.13649.

[26] A. Shrestha, "System and method for generating predictive insurance carrier appetite intelligence by observing and analyzing real-time quote submission outcomes from automated browser agent interactions with insurance carrier web portals," US Provisional Patent Application, QuoteSweep, 2025.

[27] ——, "System and method for automated multi-carrier commercial insurance quoting using parallel AI browser agents operating on authenticated insurance carrier web portals without requiring application programming interface (API) partnerships," US Provisional Patent Application, QuoteSweep, 2025.

[28] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008.

[29] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in *Proc. NeurIPS*, 2023, arXiv:2302.04761.

[30] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, and J. Tang, "AgentBench: Evaluating LLMs as agents," in *Proc. International Conference on Learning Representations (ICLR)*, 2024, arXiv:2308.03688.

[31] Catalyit, "2024 state of tech in independent insurance agencies report," Catalyit, Tech. Rep., 2024, available: https://catalyit.com/state-of-tech.

[32] Agentero, "AI Appetite Checker launch," PR Newswire, Nov. 2025.

[33] N. Halioua, A. Bloch, and A. Guez, "MARIA: A multi-agent regulatory intelligence architecture," Cleo Labs, Tech. Rep., Feb. 2026.

[34] A. Shrestha, "The QuoteSweep stated-appetite corpus, v1.0," 2026, 509 carriers; 2,031 line-of-business rows; scrape window 2026-03-23 to 2026-04-06. [Online]. Available: https://doi.org/10.5281/zenodo.20280436

[35] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[36] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.