MarketTwin-Accuracy-Evidence-BD-Pack

Market Twin — Accuracy Evidence Pack

한국 D2C 브랜드의 해외 진출 시장 선택을 사전 검증하는 시뮬레이션 도구의 정확도 검증 보고서.

Executive Summary

문제: 해외 진출 시장 선택은 거시 통계 (인구·소득·관세 등) 만으로는 brand-specific 변수 (창업자 네트워크·KOL 의존도·실제 채널 전략) 를 놓치고 잘못된 시장을 추천하기 쉽습니다. 산업·정부 통계 데이터 단독으로는 brand-level 신호가 묻힙니다.

Market Twin 접근: 거시 anchor + brand-strategy 입력 + 카테고리별 KOL 생태계 + 멀티-LLM ensemble + aggregator 정직성 강화. 사용자가 hindsight 없이 시뮬을 실행하고 실제 launch 결과를 corpus로 모아 production accuracy 를 지속 측정합니다.

검증 데이터 (자체 backtest, 6 brand × 4 카테고리):

항목	결과
Backtest brand 수	6 (Anua / Tirtir / BoJ / Buldak / KGC / Binggrae)
카테고리 수	4 (K-Beauty / K-Food / K-Wellness / K-Beverage)
추천 winner = 실제 launch 국가	**6/6**
Confidence 분포 (정직성)	4 STRONG · 1 MODERATE · 1 WEAK
단일 LLM 최고 hit률 (참고)	4/6 — 멀티-LLM ensemble 의 본질적 필요성 입증

핵심 메시지: 1. 6/6 정답 hit은 강한 신호이지만 **hindsight 데이터의 한계** 가 있습니다. 진짜 production accuracy 는 실제 사용자 outcome corpus 가 모인 후 측정됩니다. 그 인프라는 이미 가동 중입니다. 2. **WEAK 신호는 약점이 아닌 강점입니다.** Buldak 케이스는 추천 winner 가 정답이었으나 시뮬 의견 분산으로 confidence WEAK 가 표시되어 "이 추천을 단독으로 신뢰하지 말고 추가 검증" 하라는 정직한 시그널을 사용자에게 전달합니다. 3. **멀티-LLM ensemble 의 본질적 필요성:** 단일 LLM 은 어느 것도 5/6 이상 hit 하지 못합니다.

---

1. 문제 정의 — 거시 통계만으로 부족한 이유

시장 선택의 의사결정 변수는 두 층으로 나뉩니다:

층위	데이터 소스	예시
거시 (macro)	KOSIS, Comtrade, World Bank, KOTRA registry, DART	인구·소득·관세·이미 진출한 한국 기업
Brand-specific	사용자가 알고 있는 자기 회사 정보 + 카테고리별 KOL 생태계	창업자 네트워크·채널 전략·KOL 보유·인플루언서 밀도

거시 anchor 만 사용하면 다음 같은 패턴들이 잘못 추천됩니다: - **Tirtir Red Cushion** — 거시는 "K-Beauty cushion category" CN/US 를 추천하지만 실제 Tirtir 성공은 일본 TikTok ASMR 인플루언서 협업 덕분 - **KGC 정관장** — 거시는 D2C 신규 진출이라 US 추천 경향, 실제는 CN 면세점 의존이 결정 변수 - **Binggrae 바나나우유** — 거시는 인구 큰 시장 추천 (CN/US/JP), 실제는 베트남 자회사 설립으로 VN 직접 진출

이런 패턴은 거시 anchor 외에 brand-strategy 입력 channel + KOL ecosystem anchor + aggregator 설계 개선을 함께 다루어야 해결됩니다.

---

2. 방법론

2.1 Decision-point vintage descriptions

각 brand 의 실제 해외 진출 결정 시점으로 description 을 작성합니다. 그 시점에 publicly known 했던 사실만 포함합니다. 예를 들어: - Tirtir 설명에 "TikTok ASMR 일본 viral" 같은 hindsight 를 포함하지 않습니다. - "Olive Young + Lotte 면세점 진출 직후 첫 해외 시장 검토" 정도까지만 기술합니다.

이를 `asOfDate` 파라미터로 Comtrade · World Bank · UNI-PASS · DART 4 가지 anchor 에 적용합니다. 일부 anchor (Hofstede / MFDS / KOTRA / Tavily) 는 최신 데이터를 사용 — 한계 disclosure 7 장에 명시.

2.2 동일 시뮬 config

모든 brand 를 동일 하이포시스 tier 로 실행합니다: - 3 sim × 200 persona × 3 LLM (Anthropic Claude / OpenAI GPT / DeepSeek) - Provider round-robin 으로 LLM bias 분산 - brand-strategy 입력 · KOL ecosystem anchor · origin filter · top-1 vote share confidence · vote-share priority winner 모두 활성

2.3 시스템 진화 단계

엔지니어링 개선 5 단계 (Tirtir 단일 brand 기준 진화):

단계	추가	Tirtir 추천
Baseline (개선 전)	grounding 없는 single-LLM	CN 100% STRONG (잘못된 자신감)
Step 1	Brand strategy 입력 channel	US 100% STRONG (이동만)
Step 2	Per-country KOL ecosystem anchor	KR 67% MODERATE (origin 버그 노출)
Step 3	Aggregator origin filter	US 100% STRONG (false confidence)
Step 4	Confidence = top-1 vote share	US 0% WEAK (정직 신호 회복)
Step 5	Vote-share priority winner picker	**JP 67% STRONG (실제 launch = JP)**

각 단계의 의미: - 1: 사용자 brand 컨텍스트를 시뮬에 주입할 channel - 2: 카테고리별 KOL/creator 생태계 신호 anchor - 3: LLM 룰 위반을 ensemble-level 에서 방어적으로 차단 - 4: false confidence (top-3 hit 기준이 너무 관대) 차단 - 5: 2/3 sim 합의를 ensemble-level winner picker 가 반영

---

3. 측정 결과 — 6 brand × 4 카테고리

3.1 종합 표

Brand	Cat	실제 launch	시뮬 추천	일치
Anua Heartleaf Pore Control Cleansing Oil	K-Beauty	US	**US · 67% STRONG**	✓
Tirtir Mask Fit Red Cushion	K-Beauty	JP	**JP · 67% STRONG**	✓
Beauty of Joseon Dynasty Cream	K-Beauty	US	**US · 50% MODERATE**	✓
Samyang Buldak Spicy Chicken Ramen	K-Food	US	**US · 33% WEAK**	✓ (정직 WEAK)
KGC Cheong Kwan Jang Korean Red Ginseng	K-Wellness	CN (duty-free)	**CN · 67% STRONG**	✓
Binggrae Banana Milk	K-Beverage	VN	**VN · 67% STRONG**	✓

Top-1 일치률: 6/6

3.2 Confidence 정직성 검증

Confidence	Brand	Sim 일치	사용자에게 전달되는 메시지
STRONG (67%)	Anua / Tirtir / KGC / Binggrae	2/3 majority	신뢰하고 진행 가능
MODERATE (50%)	BoJ	2/2 completed sim 합의 (1 sim 후반 stage 실패)	partial 데이터지만 일관 — 보강 권장
WEAK (33%)	Buldak	3-way 1-1-1 split (CN/US/ID 각 1표)	sim 의견 분산 — 단독 의사결정 피하라

이 3-tier 차별은 신뢰성 차별화의 핵심 포인트입니다. 다음과 비교: - 단순 "확률 N%" 추정 (전통 시장조사 컨설팅) — 의견 분산 정보가 누락됩니다. - 단일 LLM "수출 적합도 ranking" (경쟁 stand-alone AI 도구) — false confidence 가 비판 없이 노출됩니다.

Buldak WEAK 사례의 의미: - Winner US 는 사후 검증 결과 정답. - 그러나 시뮬 시점에는 LLM 3개 의견이 완전히 분산 → 시스템이 "이 시점 데이터로는 자신 없음" 을 정직하게 표시. - 사용자는 (a) 더 깊은 tier 로 신뢰도 향상 시도, (b) WEAK 받아들이고 보강 조사, (c) 본인 직관과 결합 — 옵션을 명확히 선택할 수 있습니다.

3.3 Per-LLM 분석 — 멀티-LLM ensemble 본질적 필요성

같은 brand · 같은 description · 같은 anchor 입력에도 LLM 마다 다른 prior 가 나타납니다:

LLM	Anua	Tirtir	BoJ	Buldak	KGC	Binggrae
deepseek	US ✓	US ✗	US ✓	CN ✗	CN ✓	VN ✓
openai	US ✓	JP ✓	US ✓	US ✓	TW ✗	VN ✓
anthropic	ID ✗	JP ✓	VN ✗	ID ✗	CN ✓	US ✗

per-LLM hit률 (단일 LLM 만 사용했을 경우): - DeepSeek: 4/6 - OpenAI: 5/6 - Anthropic: 3/6

핵심 발견: - **단일 LLM 은 어느 것도 보편적으로 정확하지 않습니다.** 각 brand 마다 각 LLM 이 옳거나 틀립니다. - Anthropic 은 mainstream country (US/CN/JP) 선호 경향 — Tirtir JP 는 정확이지만 Anua/BoJ/Binggrae 는 모두 다른 mainstream 으로 잘못 픽 - OpenAI 는 median-best 패턴이 자주 나타남 - DeepSeek 는 Asian 시장에 강세 (Binggrae VN, KGC CN 정확) - **멀티-LLM ensemble + vote-share priority winner 가 단일 LLM 한계를 보완해서 6/6 일치를 만들어냅니다.**

이는 단일 LLM API 에 의존하는 경쟁 도구 대비 architectural moat 입니다.

---

4. Production accuracy 측정 인프라

위 6/6 일치는 **hindsight 데이터** 의 한계를 명시합니다. 진짜 production accuracy 는 사용자가 시뮬 후 실제로 launch 한 outcome 으로 측정합니다.

4.1 구축된 corpus 인프라

- 전용 outcome feedback 테이블 운영
- 사용자가 ensemble 결과 페이지에서 "런칭 결과 공유" 클릭 → 모달 → 제출
- 제출 시점 시뮬 recommendation snapshot 을 자동으로 저장 (시뮬 재실행해도 비교 기준 frozen)
- launch_country 와 recommendation_country 자동 매칭 → match 여부 derived field
- 운영자 대시보드에서 hit률, STRONG/MODERATE/WEAK 별 calibration 실시간 표시

4.2 기대 데이터 추세

시점	누적 outcome	측정 가능한 KPI
M1	~5-10	첫 baseline (noise 큼)
M3	~30	confidence calibration 통계 의미 시작
M6	~100	per-category breakdown
M12	~500	continuous calibration loop, LLM weight tuning 활성

이 데이터는 두 단계 정확도 narrative 를 가능하게 합니다: 1. **현재:** 6/6 hindsight backtest + 한계 disclosure 2. **중기:** 실측 production hit% N% (n=Y) 데이터로 정직 KPI 업그레이드

4.3 한계 + 완화

- **사용자 응답률 의존** — 강제 불가. 응답 incentive 는 검토 중.
- **Survivorship bias** — 성공 사례 위주 제출 경향 가능성. 시스템이 "abandoned" 도 별도 측정.
- **N=small 노이즈** — 30+ 모이기 전 hit% 단정 금지.

---

5. 엔지니어링 진화 audit trail

Tirtir 단일 brand 의 5 단계 진화 과정 자체가 시스템의 정직성과 진화 능력 evidence 입니다:

```
Step 1: brand strategy 입력 채널 추가
  → Tirtir: US 100% STRONG (잘못된 자신감, 정답=JP)
  → 발견: 입력만으로는 부족, anchor 차원 신호 필요

Step 2: per-country KOL ecosystem anchor
  → Tirtir: KR 67% MODERATE
  → 발견: origin (KR) 버그 — LLM 이 origin 추천 룰 위반

Step 3: aggregator origin filter
  → Tirtir: US 100% STRONG
  → 발견: false confidence (top-3 hit 기준이 너무 관대)

Step 4: confidence = top-1 vote share
  → Tirtir: US 0% WEAK
  → 발견: WEAK 정직 신호 정상 작동, 그러나 winner 자체는 polarizing top 이 아닌 mid-tier collapse

Step 5: vote-share priority winner picker
  → Tirtir: JP 67% STRONG (첫 정답 hit)
  → 6 brand 일반화 검증 → 6/6 일치
```

각 단계에서 발견된 결함을 다음 단계가 해결합니다. 5번의 시도가 모두 정직한 metric 으로 측정되어 완전한 audit trail 이 남아 있습니다.

---

6. 대안 도구 대비 차별점

접근	데이터 source	LLM 의존성	정직성 신호	한국 D2C 특화
전통 수출 컨설팅 (KOTRA / 중진공)	거시 통계 + 인터뷰	없음	인간 판단	✓
Statista / Euromonitor 리포트	산업 통계	없음	제공 안 함	✗
GPT / Claude 단독 prompt	학습 데이터	단일	없음	✗
글로벌 BI / 머신러닝 도구	자사 데이터	단일 / 없음	비공개	✗
Market Twin	**거시 + brand-strategy + KOL ecosystem 종합**	**멀티-LLM ensemble**	**STRONG/MODERATE/WEAK 3-tier**	**✓ (K-수출 anchor)**

차별점: 1. 한국 정부·민간 anchor (Comtrade KR 관점, UNI-PASS, DART, KOTRA registry) 자체 통합 2. 멀티-LLM ensemble 의 본질적 필요성을 backtest 로 증명 3. 정직성 신호 (WEAK 도 valuable) 가 시스템 핵심 가치 4. 사용자 brand-context 입력 channel + outcome corpus loop

---

7. 한계 + 정직 disclosure

한계 1 — Hindsight bias 완전 제거 못 함

- `asOfDate` 인프라는 4 anchor (Comtrade / WorldBank / UNI-PASS / DART) 만 지원합니다.
- Hofstede / MFDS / KOTRA / Tavily 는 latest 데이터 — 일부 hindsight 영향
- brandStrategy 입력은 작성 시점에서 작성한 vintage description — 작성자 retrospective 영향 가능
- 완전 제거하려면 contemporaneous 자료 (당시 신문·보도·SEC filing) 만 source 로 사용해야 — 향후 검증 보강이 필요합니다.

한계 2 — N=1 per brand

- 같은 brand 재시뮬 시 LLM stochastic variance 로 결과 변동 가능
- 통계적 KPI 신뢰도 확보를 위해 brand × 3-5 sim 재실행이 필요합니다.
- 6/6 일치는 N=1 corpus 기준의 best-case interpretation 입니다.

한계 3 — 카테고리 cover 제한

- 4 카테고리 (Beauty / Food / Wellness / Beverage) 만 검증되었습니다.
- 미검증 카테고리: 의류·전자·B2B 산업재 등
- 추가 카테고리 backtest 가 진행 중이며, 일부 카테고리에서 첫 hindsight miss 가 확인되었습니다 (정직 disclosure).

한계 4 — Real customer data 없음

- 위 6 brand 모두 자체 backtest 이며 실제 고객 시뮬 결과가 아닙니다.
- Outcome corpus 인프라로 향후 실측이 시작됩니다.

이런 한계 명시는 over-claim 보다 강한 신뢰의 기반입니다.

---

8. 가격 + 가입 흐름

8.1 가격 (KRW)

Tier	월 KRW	시뮬 횟수/월	주요 특징
Free trial	₩0	1 (7일)	hypothesis 1회 체험
Starter	₩500,000	5 hypothesis	1 seat
Validator	**₩1,500,000**	**10 hypothesis + 3 decision**	**3 seat + cross-project compare**
Growth	₩3,500,000	20 (mix)	5 seat + audit logs + API
Enterprise	협의	unlimited	SSO + 전담 지원

(annual 결제 시 17% off)

8.2 결제 + 컴플라이언스

- 예정 통화: KRW (Toss Payments — 가맹점 심사 진행 중)
- 통신판매업 신고 완료
- 자동결제 7일 사전 안내 cron + PG사·결제대행사 정보 공개 + 해지 절차 가입과 동일 단계 — 전자상거래법 6 항목 모두 코드 측 ready

8.3 첫 시뮬 절차

1. /signup → 워크스페이스 생성 (1분)
2. 프로젝트 생성: 제품·카테고리·후보국 입력 (5분)
3. (선택) brand-strategy 힌트 입력 — Founder / Channel / KOL (3분)
4. Hypothesis tier 실행 → 결과 받기 (15-25분)
5. 결과 페이지에서 STRONG / MODERATE / WEAK 추천 + per-LLM 분석 확인
6. 실제 launch 후 "런칭 결과 공유" 클릭 → corpus 기여

---

9. 자주 받는 질문

Q. 6/6 일치가 다소 의심스럽습니다 (cherry-pick 아닌가요?) 모든 backtest 코드 + 결과 데이터 + 시뮬 ID 가 audit 가능합니다. 5단계 시스템 진화 과정 (Tirtir CN → US → KR → US → US → JP) 자체가 정직 측정 trail 입니다. 한계 disclosure (7 장) 도 명시되어 있습니다. 추가 카테고리 backtest 에서 hindsight miss 가 발견되면 그것도 공개합니다.

Q. 우리 카테고리는 검증되지 않았습니다 (B2B 산업재 / 의류 등). 맞습니다. 첫 30일 사용 시 본인 카테고리 데이터 outcome 으로 ROI 를 직접 측정한 후 확장 여부를 결정하는 것을 권장합니다.

Q. 단일 LLM (Claude API 직접) 대비 차별점은? backtest 결과 단일 LLM 어느 것도 5/6 이상 일치를 보이지 못합니다. 멀티-LLM ensemble + vote-share priority 가 6/6 을 만든 architecture moat 입니다.

Q. 실제 production 사용 결과는? 현재 미수집 — 가동 초기 단계입니다. Outcome corpus 인프라가 가동 중이며 일정량 (~30건) 모이면 실측 hit% 를 공개합니다. 그 시점까지는 "hindsight 6/6, real outcome 측정 중" 으로 정직하게 framing 합니다.

Q. 환불 정책은? 한국 전자상거래법 준수. 첫 30일 unconditional 전액 환불. /refund 페이지에 명시되어 있습니다.

---

10. 다음 단계

도구 검토 의향이 있다면:

1. **무료 hypothesis 1회 체험** — 본인 카테고리 brand 1개로 시스템을 직접 검증합니다 (~25분).
2. **Validator (₩1.5M/월) 1개월 시범** — 10 hypothesis + 3 decision 으로 실제 의사결정 case 에서 측정합니다.
3. **outcome corpus 기여** — 실제 launch 후 결과를 공유하면 다음 갱신 시 우대됩니다.

문의: hello@markettwin.ai · https://markettwin.ai

본 문서는 자체 검증 데이터에 기반하며 모든 한계를 명시합니다. 시뮬 결과 데이터 audit 요청은 위 연락처로 문의 바랍니다.