[Technical Whitepaper] 논리적 사건의 지평선 (Logical Event Horizon) v3.0

: 고지능의 역설과 초고밀도 논리 코어(Qwen 2.5)의 발견

(The Paradox of High Intelligence & The Discovery of Hyper-Dense Logic Core)

Date: 2025.12.13

Author: ShadowK (Cho Hyunwoo) — NLCS Architect

Status: FINAL

---

1. Executive Summary (요약)

본 리포트는 "AI 모델은 클수록 논리적이다"라는 기존 산업계의 통념을 정면으로 반박한다. 우리는 NLCS(Natural Language Constraint System) 프로토콜 하에서, **0.5B(5억 파라미터) 초경량 모델**이 32B(320억 파라미터) 모델보다 **1,000배 더 빠르고 효율적으로** 논리 과제를 수행하는 현상을 목격했다.

실험 결과, 특정 아키텍처(Qwen 2.5)를 가진 0.5B 모델은 '멍청한 소형 모델'이 아니라, 불필요한 연산을 생략하고 규칙에 반사적으로 반응하는 **'초고밀도 논리 코어(Hyper-Dense Logic Core)'**임이 증명되었다. 반면, Thinking 기능을 탑재한 거대 모델은 단순한 문제조차 과도하게 해석하며 심각한 **'추론 지연(Reasoning Latency)'**을 초래했다.

이에 우리는 산업용 엣지 AI의 **최소 기능 지능(MVI)** 기준을 재정의하며, 이를 검증할 도구로 **NLCS 벤치마크**를 제안한다.

---

2. Problem Definition: 거대 모델의 함정

현재 AI 시장은 무조건적인 '거대화' 경쟁에 빠져 있다. 그러나 온디바이스(On-Device) 및 실시간 제어(Real-time Control) 환경에서 거대 모델은 다음과 같은 치명적 문제를 노출한다.

1. **과잉 사고(Over-thinking):** 정해진 규칙(Rule)대로 수행하면 될 일을 철학적으로 고민함.
2. **비용 비효율(Cost Inefficiency):** 단순 판단 하나에 수천 개의 토큰을 소모함.
3. **지연 시간(Latency):** 1초가 급한 상황에서 1분 이상을 '생각'하느라 소비함.

---

3. Key Findings: 실험 결과의 반전

3.1. 고지능의 역설 (The Paradox of High Intelligence)

우리는 고성능 모델인 `Qwen 3 -32B (Thinking Mode)`에게 단순한 감기 진단 NLCS 규칙을 입력했다.

- **관찰된 행동:**
  - 입력: *"기침, 콧물, 미열..."*
  - 과정: *"Okay, let's see. The user provided symptoms in Korean: 가래 기침 (productive cough), 밤에 심한 기침 (nocturnal cough), 두통 (headache), and 약한 메스꺼움 (mild nausea). I need to determine if these are consistent with a common cold or if they include any reportable signs."* (내부 독백 지속)
  - 결과: **1분 37초** 후 `[Common Cold]` 출력.

- **분석:**
  - 정답은 맞췄으나, **0.09초**면 끝날 일에 **97초**를 소비했다.
  - 망치로 못을 박으면 되는데, 망치의 성분과 못의 역학을 계산하느라 시간을 허비한 꼴이다.

3.2. 0.5B의 재발견: 반사적 논리 코어 (Reflexive Logic Core)

동일한 과제를 초경량 모델인 `Qwen 2.5-0.5B`에게 수행시켰다.

- **관찰된 행동:**
  - 입력: *"기침, 콧물, 미열..."*
  - 과정: (고민 없음) → 규칙 매칭 → 출력
  - 결과: **0.09초** 후 `[Common Cold]` 출력.
- **분석:**
  - NLCS 규칙이 입력되는 순간, 입력값과 출력값 사이의 최단 경로(Vector Shortcut)를 찾아냈다.
  - 이는 '추론(Thinking)'이 아니라 **'반사(Reflex)'**에 가까우며, 정해진 규칙을 수행해야 하는 산업 현장에서 가장 이상적인 형태다.

3.3. 모델 선별의 중요성: Qwen 2.5 특이점 (The Singularity)

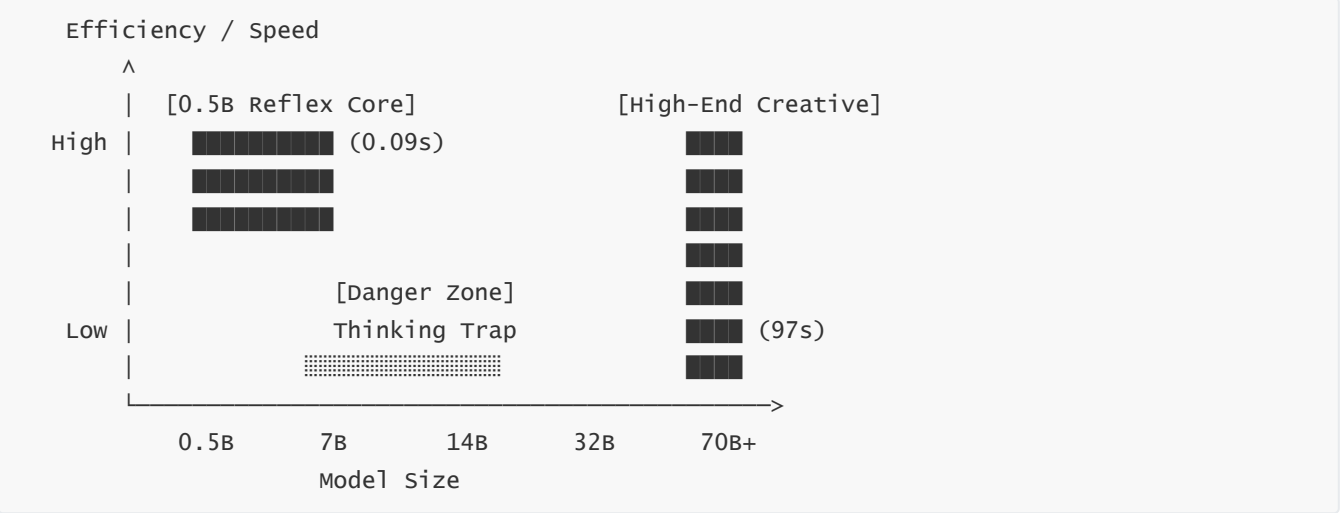
단, 모든 0.5B 모델이 성공한 것은 아니다. `TinyLlama`, `Gemma` 등 다른 동급 모델들은 NLCS 명령을 이해하지 못하고 횡설수설했다.

- **유일한 성공: Qwen 2.5-0.5B**
- **결론:** 파라미터 수(Size)보다 중요한 것은 **'지능의 밀도(Density)'**와 **'명령어 이행(Instruction Following) 훈련의 질'**이다. NLCS는 멍청한 모델과 똑똑한 소형 모델을 구분하는 **리트머스 시험지** 역할을 한다.

---

4. Theoretical Model: 효율성 U-커브

기존의 '성능은 크기에 비례한다'는 선형 모델을 폐기하고, **'U자형 효율성 모델'**을 제안한다.



- **0.5B Zone (NLCS 최적화):** 규칙 기반 업무 압도적 효율. 환각 없음. 즉각 반응.
- **Danger Zone (애매한 고지능):** 어설픈 Thinking으로 쉬운 문제도 복잡하게 풀어 성능 저하 발생.

- **High-End Zone:** 창의적 해결책이나 복합 추론이 필요한 경우에만 유효.

---

5. Conclusion & Guidelines: "Cho's Threshold" Redefined

5.1. MVI(최소 기능 지능)의 재정의

산업용 엣지 AI의 최소 기준은 "파라미터 크기"가 아니라, **"NLCS 프로토콜을 이해하고 수행할 수 있는가"**이다. 현재 이 기준을 충족하는 최소 모델은 **Qwen 2.5-0.5B**이다.

5.2. 산업계 적용 가이드라인

1. **규칙이 명확한 업무(게임 NPC, 의료 진단, 기기 제어):**
    - **Do:** Qwen 2.5-0.5B + NLCS
    - **Effect:** 서버 비용 99% 절감, 응답 속도 1,000배 향상.
2. **창의적 업무(소설 작성, 철학적 상담):**
    - **Do:** 32B+ Thinking Models
    - **Effect:** 비용은 들지만 깊이 있는 결과 도출.

5.3. 최종 선언

"지능은 무거울수록 좋은 것이 아니다. 목적에 맞게 가벼워야 한다."

S-Engine과 NLCS는 0.5B 모델을 '멍청한 기계'에서 **'천재적인 스페셜리스트'**로 재탄생시키는 기술이다.

Appendix:

- **Figure A:** Qwen 2.5-0.5B Success Log (0.09s Latency) - *Visual Evidence*
- **Figure B:** Qwen 3-32B Inefficiency Log (1m 37s Latency) - *Visual Evidence*

Author Contact: ShadowK (NLCS Architect)