홈랩 구축 체크리스트

로컬 LLM을 세우고, 어디서나 안전하게 잇기까지 20단계

01 하드웨어와 모델 고르기

- [ ] VRAM 예산 확인: 7B는 4~5GB, 13B는 8~10GB, 70B는 40~48GB
- [ ] 모델 전체가 VRAM에 올라가는지 확인 — 넘치면 수십 배 느려진다
- [ ] 입문은 16GB에 Gemma 3 9B 또는 Llama 3.2 8B부터
- [ ] Apple Silicon은 통합 메모리 기준 — M3 Pro 18GB면 13B급
- [ ] 한국어는 Qwen 2.5 계열, 코딩은 Qwen 2.5 Coder 32B

02 로컬 LLM 세우기

- [ ] 설치 스크립트 실행 후 ollama --version 으로 확인
- [ ] ollama run 으로 첫 모델을 받고 토큰/초 체감 속도를 잰다
- [ ] Open WebUI를 Docker Compose로 올린다 (3000:8080)
- [ ] OLLAMA_BASE_URL과 host.docker.internal 게이트웨이 설정
- [ ] 아레나 모드로 후보 모델 2~3개를 같은 프롬프트로 비교한다

03 Tailscale로 잇기

- [ ] 설치 스크립트 후 tailscale up 으로 SSO 인증
- [ ] 둘째 기기에서 같은 계정으로 up — 100.x 대역 할당 확인
- [ ] Admin Console에서 MagicDNS를 켜고 기기명으로 접속한다
- [ ] systemctl enable --now tailscaled 로 재부팅 뒤에도 유지
- [ ] 서버·컨테이너는 브라우저 로그인이 안 되니 인증 키를 쓴다

04 열기 전에 잠그기

- [ ] LLM 서버를 공인 IP로 직접 열지 않는다 — 기본 인증이 없다
- [ ] Open WebUI는 WEBUI_AUTH=true 로 계정과 기록을 분리한다
- [ ] 공개가 필요하면 리버스 프록시 + TLS + 기본 인증을 앞에 둔다
- [ ] ACL로 접근 범위를 좁힌다 — src 그룹에서 dst 태그:포트로
- [ ] SSH는 ACL의 ssh 섹션에서 사용자와 대상을 따로 제한한다

전체 가이드 2편은 블로그의 practice 트랙에 있다 — 로컬 LLM 구축 · Tailscale 실전 구성