

۱۶۹ بنچمارک استاندارد غربالگری نظام‌مند ۲۰۱۵–۲۰۲۶	۷ مدالیت‌های حس-حرکتی متن، تصویر، صوت، ویدیو، 3D، سنسور، عمل	۹ بُعد قابلیت شناختی ۳ لایه ادراک، شناخت و عاملیت فیزیکی	۷ عنصر ماده تاریک (Dm) ۷ موضوع بکر پایان‌نامه ارشد و دکتری
--	---	---	---

۱. کالبدشکافی تاکسونومی دوگانه (۷ مدالیت‌های حس-حرکتی x ۹ بُعد قابلیت شناختی در ۳ لایه)

لایه شناختی	ابعاد قابلیت (Capability Dimensions)	مدالیت‌های ورودی/خروجی حس	فرمول شاخص پوشش عمق قابلیت (Coverage Depth)
لایه ۱: ادراک و مکان‌یابی Perception & Grounding	- CAP_PERC ادراک و تشخیص وجود - CAP_GROUND مکان‌یابی و زمینه‌سازی (جعبه، زمان)	- MOD_TEXT متن ساخت‌یافته و پرامپت - MOD_IMAGE تصویر دوبعدی ایستا - MOD_AUDIO صوت خام، گفتار و اکوستیک - MOD_VIDEO استریم ویدیویی پیوسته - MOD_3D ابرنقاط و مش سه‌بعدی - MOD_SENSOR سنسورهای LiDAR/IMU/Tactile - MOD_ACTION کنترل موتور و گشتاور بازو	$c(b) = (1/9) \sum_{k=1..9} [cap_k(b) = 1]$ شاخص متعادل پوشش: یک بنچمارک تک‌بعدی مثل AudioCaps نمره $c=1/9$ می‌گیرد، در حالی که بنچمارک عاملیتی-رباطیک پیشرفته مثل ALFRED نمره $c=6/9$ (۶ بعد فعال) کسب می‌کند.
لایه ۲: مدل‌سازی شناختی Cognitive Modeling	- CAP_REASON استدلال چندگامی و منطقی - CAP_INSTR تبعیت از دستورالعمل پیچیده - CAP_KNOW دانش تخصصی و دانشگاهی - CAP_MEM حافظه دوربرد (ساعت‌ها ویدیو)		
لایه ۳: عاملیت و تعامل فیزیکی Agency & Embodiment	- CAP_PLAN برنامه‌ریزی سلسله‌مراتبی و زنجیره‌ای - CAP_TOOL فراخوانی مفسر، پایگاه‌داده و API - CAP_EMBOD تجسم فیزیکی، حل تعارض و فیدبک		

۲. ساختار کلان جدول تناوبی ۱۶۹ بنچمارک چندوجهی در ۶ خانواده مدالیت

خانواده مدالیت	تعداد	نمادهای معرف	بازه سال	میانگین c(b)	ویژگی‌های کلیدی ارزیابی و نماینده‌های شاخص
Vision-Language	n=47	MMB, MME, MMMU, POPE	2015–2025	0.34	تمرکز شدید بر ادراک و استدلال تصویری ایستا، اشباع سریع و نشت داده در پیش‌آموزش
Document & Diagrams	n=8	DocVQA, ChartQA, PlotQA	2020–2024	0.31	استخراج ساختار جداول، فهم هندسه نمودارها و چالش ترجمه بصری به کد پایتون
Video & Temporal	n=27	Video-MME, LVBench, EgoSchema	2018–2025	0.42	درک توالی زمانی رویدادها، چالش شورت‌کات تک‌فریم و مصرف سرسام‌آور کانتکست
Audio & Speech	n=24	AudioBench, VoiceAgent, Wav2SQL	2020–2025	0.38	گذار از رونویسی صرف (ASR) به درک لحن، طعنه، گوینده و فراخوانی ابزار بر پایه صوت
3D & Spatial	n=10	ScanQA, SQA3D, 3D-LLM	2022–2025	0.40	زمینه‌سازی مکانی در ابرنقاط سه‌بعدی، پاسخ به سوال موقعیت‌محور و مسیریابی ۶ درجه آزادی
Embodied & Agents	n=53	ALFRED, WebArena, DROID, RLBench	2019–2026	0.58	بیشترین عمق قابلیت، اقدام در جهان شبیه‌سازی و واقعی، تعامل ابزاری و چالش Sim-to-Real

۳. جعبه طلایی: ۷ موضوع بکر پایان‌نامه ارشد و دکتری استخراج‌شده از «ماده تاریک ارزیابی» (Dm1–Dm7)

کد عنصر	عنوان موضوع پایان‌نامه پیشنهادی	مدالیت‌ها و قابلیت‌های هدف	ایده پژوهشی و شکاف نوآورانه جهت چاپ مقاله ISI Q1
Dm1	مکالمه دیداری-شنیداری همگام بی‌درنگ	Text+Img+Audio Perc+Reason	طراحی بنچمارک سنجش قطع صحبت نابهنگام (Barge-in)، تأخیر زیر ۲۰۰ میلی‌ثانیه و حالات چهره همزمان با لحن گفتار.
Dm2	ایجنیت‌های ویدیویی ابزارمحور (Video+Tool)	Video+Action Plan+Tool (n=0)	شکاف کاملاً خالی جدول (n=0)! تدوین اولین بنچمارک که در آن مدل با دیدن جریان ویدیو، کد API بنویسد یا کنسول وب را کنترل کند.
Dm3	ارزیابی حلقه بسته حس لامسه و نیرو در رباتیک	Sensor+Action Embod+Plan	سنجش توانایی مدل در انطباق گشتاور و لغزش جسم با سنسورهای GelSight/Tactile در تسک‌های جابجایی ظریف اشیاء شکننده.
Dm4	حافظه فضایی سه‌بعدی پویا در صحنه‌های متغیر	3D+Video Mem+Ground	سنجش پایداری درک مدل از اشیایی که در غیاب ربات جابجا شده‌اند در محیط‌های مسکونی یا کارخانه‌ای در حال تغییر.
Dm5	سوزن در انبار کاه چندساعته در ویدیو و صوت	Video/Audio Long-Mem	ارزیابی بازیابی وقایع چندثانیه‌ای در استریم‌های ۵ تا ۱۰ ساعته بدون شورت‌کات‌های متنی و با کاندیداهای منفی گمراه‌کننده.
Dm6	استدلال علّی و مداخله فیزیکی پادواقعی	Image+Video Reason+Embod	ارزیابی مدل با سناریوهای خلاف واقع (What-if): «اگر این لیوان حذف شود کدام شیء سقوط می‌کند؟» برای آزمون درک قوانین فیزیک.
Dm7	بنچمارک‌های خودجست‌یافته با قناری رمزنگاری	All Modalities Integrity	تولید رویه‌ای محیط‌های آزمون یک‌بارمصرف با کلید قناری جهت جلوگیری تضمینی از نشت داده و تقلب در پیش‌آموزش مدل‌ها.

۴. گالری کالبدشکافی ۴ پاتولوژی بحرانی شکست مدل‌ها (Diagnostic Failure Modes)

پاتولوژی ۱: توهم شیء ناشی از سوگیری زبانی

بنچمارک‌های معرف: POPE, HallusionBench, MME, ScanQA
مکانیزم شکست: تقدم لایه‌های زبانی دکودر بر توکن‌های بصری؛ مدل بر اساس هم‌آیی کلمات (مثلاً چنگال در کنار بشقاب) حدس می‌زند نه دیدن واقعی شیء!

راهکار پروتکلی SoK: نمونه‌برداری منفی متعادل و خصمانه (Adversarial Splits) + راستی‌آزمایی با ماسک بخش‌بندی IoU.

پاتولوژی ۳: ناشناختی لحن در معماری آشناری صوت

بنچمارک‌های معرف: VoiceAgentBench, Wav2SQL, BFCL-Audio
مکانیزم شکست: تبدیل صوت به متن (ASR) تمام اطلاعات آهنگ گفتار (Prosody)، طعنه، اضطراب و احساسات را نابود کرده و خطای متنی به ابزار سرایت می‌کند.

راهکار پروتکلی SoK: مدل‌های سرتاسری پیوسته صوت (Speech-to-Token) + تست در نسبت‌های نويز متغیر و تأخیر تعاملی.

۵. چرخه ۵ مرحله‌ای آلودگی داده و قناری رمزنگاری

- مرحله ۱ (انتشار باز): سوالات و پاسخ‌های بنچمارک در گیت‌هاب/وب منتشر می‌شوند.
- مرحله ۲ (خرش کور): خزشگرهای وب دیتای تست را در کورپوس پیش‌آموزش ذخیره می‌کنند.
- مرحله ۳ (حفظ طوطی‌وار): مدل در مرحله Pre-training سوالات را حفظ می‌کند.
- مرحله ۴ (اشباع کاذب): نمره مدل در سال اول به بالای ۹۵٪ می‌رسد (توهم هوش).
- مرحله ۵ (سقوط میدانی): مدل در اولین تست داده‌های دنیای واقعی کاملاً شکست می‌خورد.

راهکار قناری رمزنگاری: تعبیه رشته هش یکتا (Canary GUID) داخل داده‌های تست تا خزشگرهای مسئولیت‌پذیر دیتای آزمون را حذف کنند.

۷. بخش ویژه کارگاه عملی زنده روی گوشی دانشجویان (Smartphone Live Lab — دقیقه ۳۸ تا ۴۸)

همین حالا بارکد مقابل را با دوربین گوشی خود اسکن کنید | alirezaabbasazadeh.github.io/LATEX-Sok/

تمرین عملی ۱ (شکار موضوع در جدول تناوبی): با گوشی خود وارد تب ۱ پلتفرم شوید. فیلتر مدالیت‌ها را روی Audio و قابلیت را روی Tool بگذارید و بنچمارک‌های موجود را ببینید. سپس در ماتریس حرارتی ۹×۷ به تقاطع خالی Video + Tool (n=0) دقت کنید که یک موضوع بکر پایان‌نامه است!

تمرین عملی ۲ (ژنراتور کارت پروتکل و کد LaTeX): وارد تب ۴ سایت (Protocol Card Generator) شوید. مدالیت‌ها و قابلیت‌های موضوع پایان‌نامه شخصی خود را تیک بزنید؛ سیستم در لحظه کد جدول استاندارد LaTeX مقاله شما را آماده کپی تحویل می‌دهد!

مدرس در کلاس همزمان این دو تمرین را روی پرده پروژکتور اجرا کرده و نتایج را با دانشجویان بررسی می‌نماید.

