

SEA-TauBench:

Adapting Tool-Agent-User Evaluation Into Low-Resource Southeast Asian Languages



Team 1
SeaCrowd Apprentice 2026

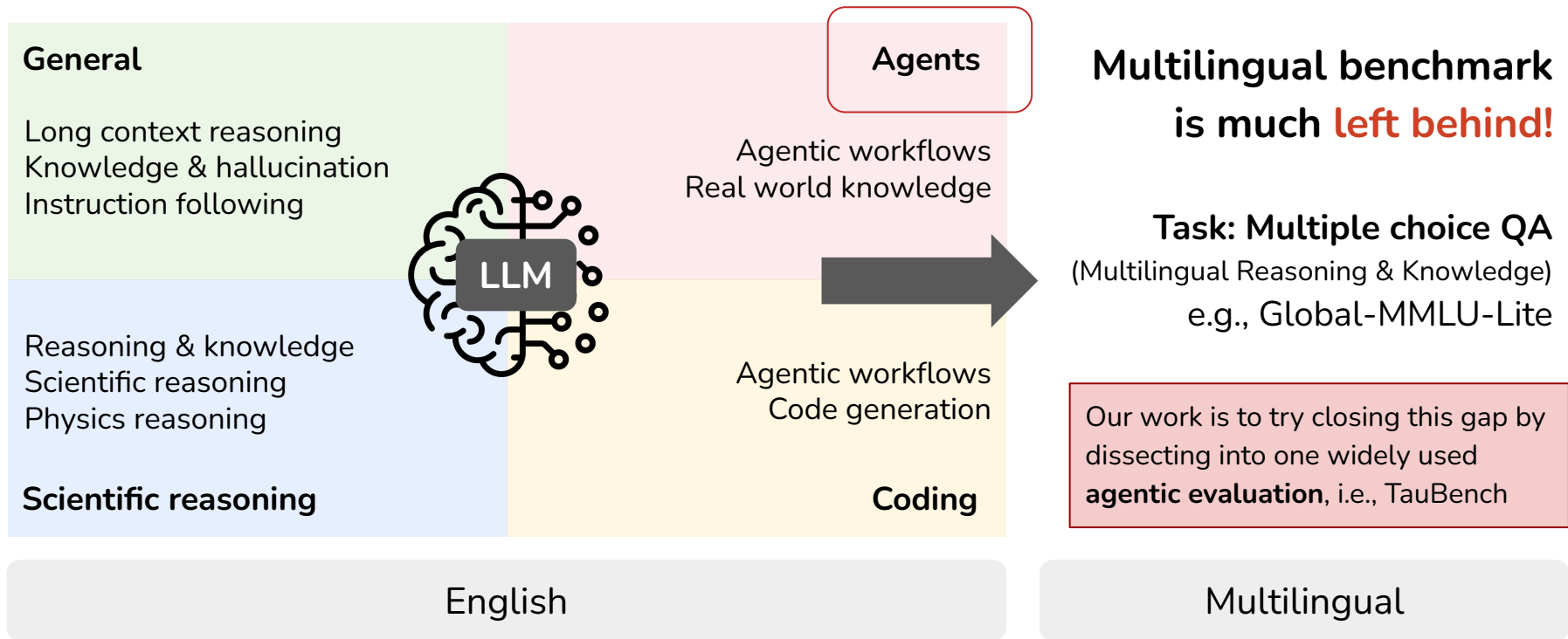


Samuel Cahyawijaya • Patomporn Payoungkhamdee

Saksorn Ruangtanusak • Kittiphath Leesombatwathana • Vissuta Gunawan Lim • Aulia Adila • My (Chiffon) Nguyen

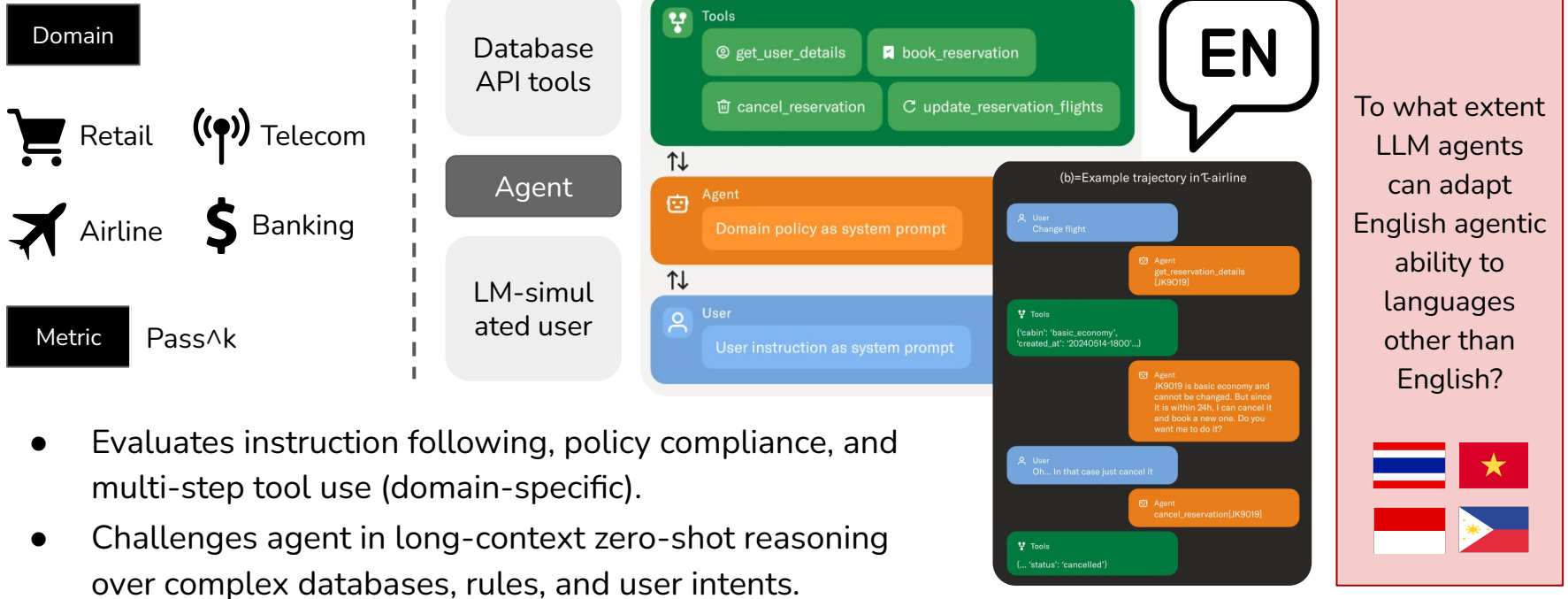
LLM capabilities today?

<https://artificialanalysis.ai/methodology/intelligence-benchmarking>



Have you heard about TauBench?

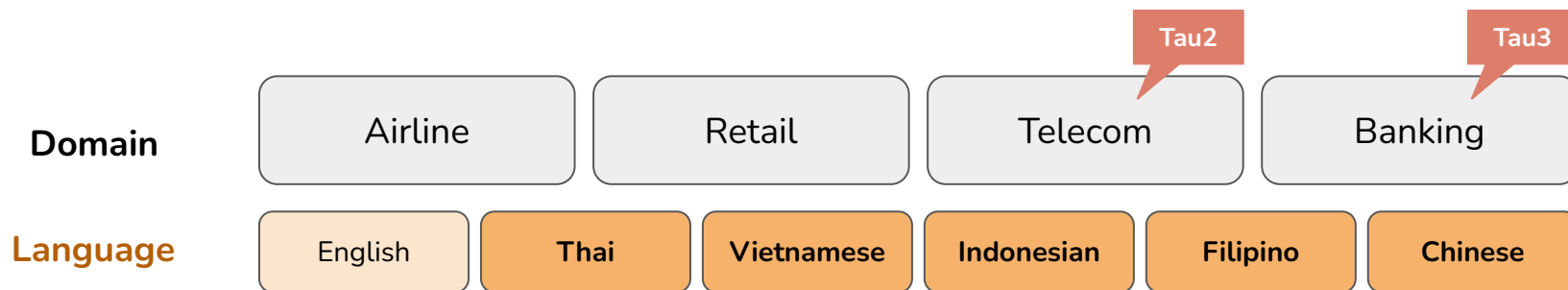
Tool-Agent-User interaction benchmark to measure agent ability in realistic task settings.



- Evaluates instruction following, policy compliance, and multi-step tool use (domain-specific).
- Challenges agent in long-context zero-shot reasoning over complex databases, rules, and user intents.

Introducing SEA-TauBench

Extending the TauBench framework to evaluate multilingual Tool-Agent-User interaction.

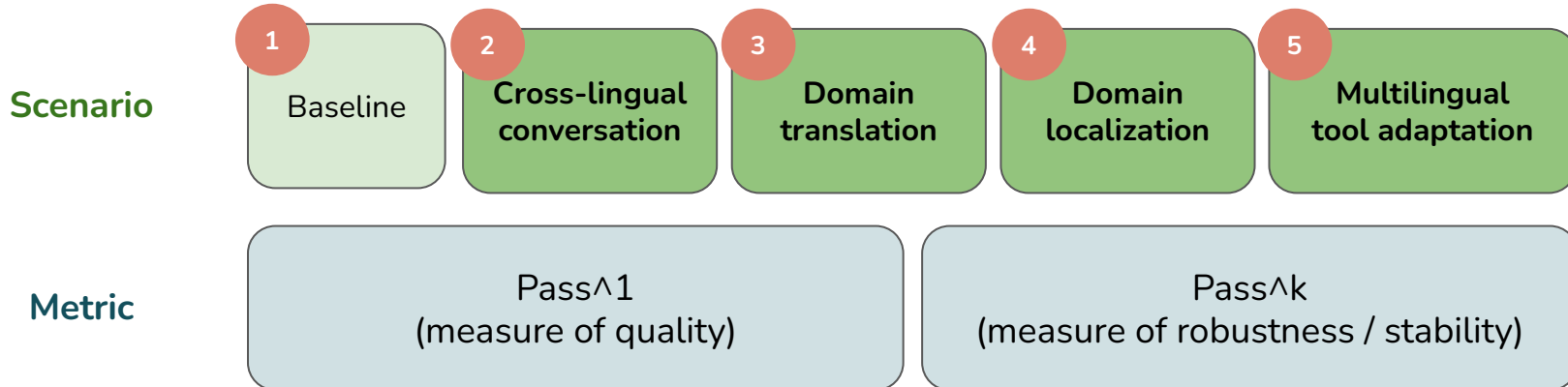


Introducing SEA-TauBench

Extending the TauBench framework to evaluate multilingual Tool-Agent-User interaction.

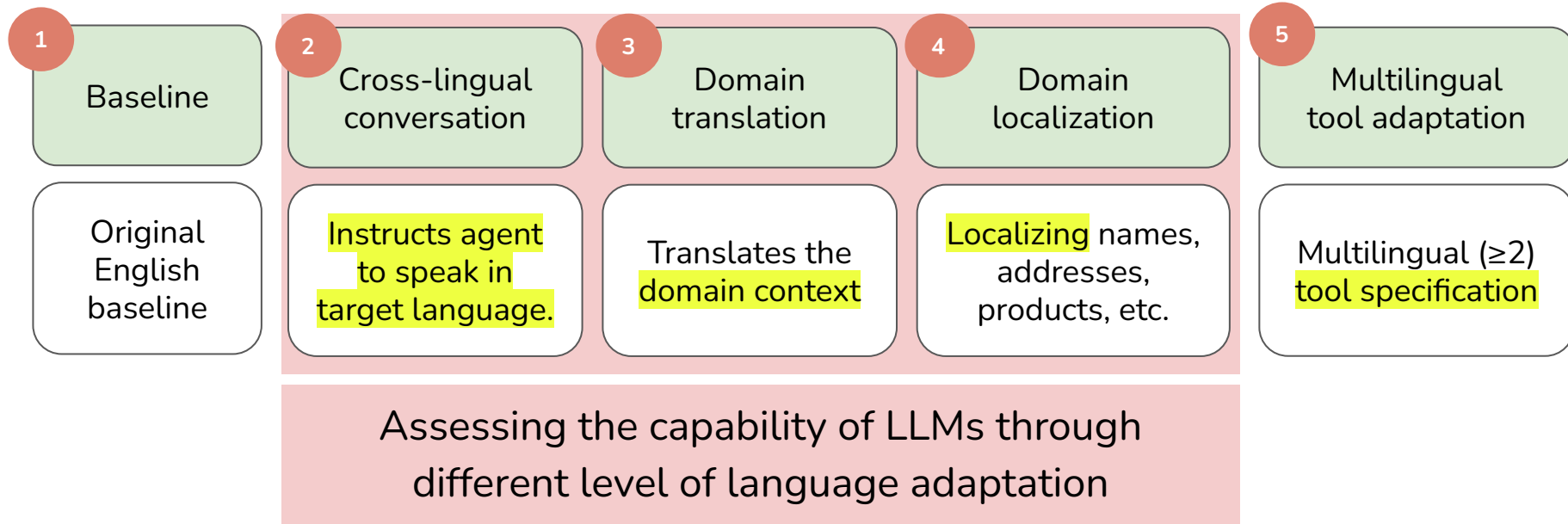
Our big question:

To what extent LLM agents can adapt English agentic ability to non-English languages?



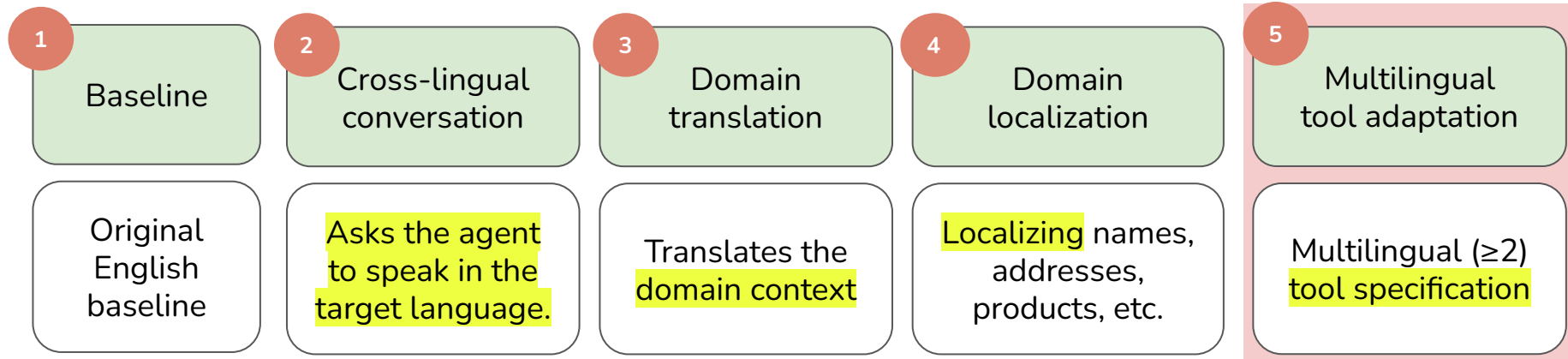
What we're curious about?

Research questions are centered around the agent adaptability in cross- and multilingual environments.



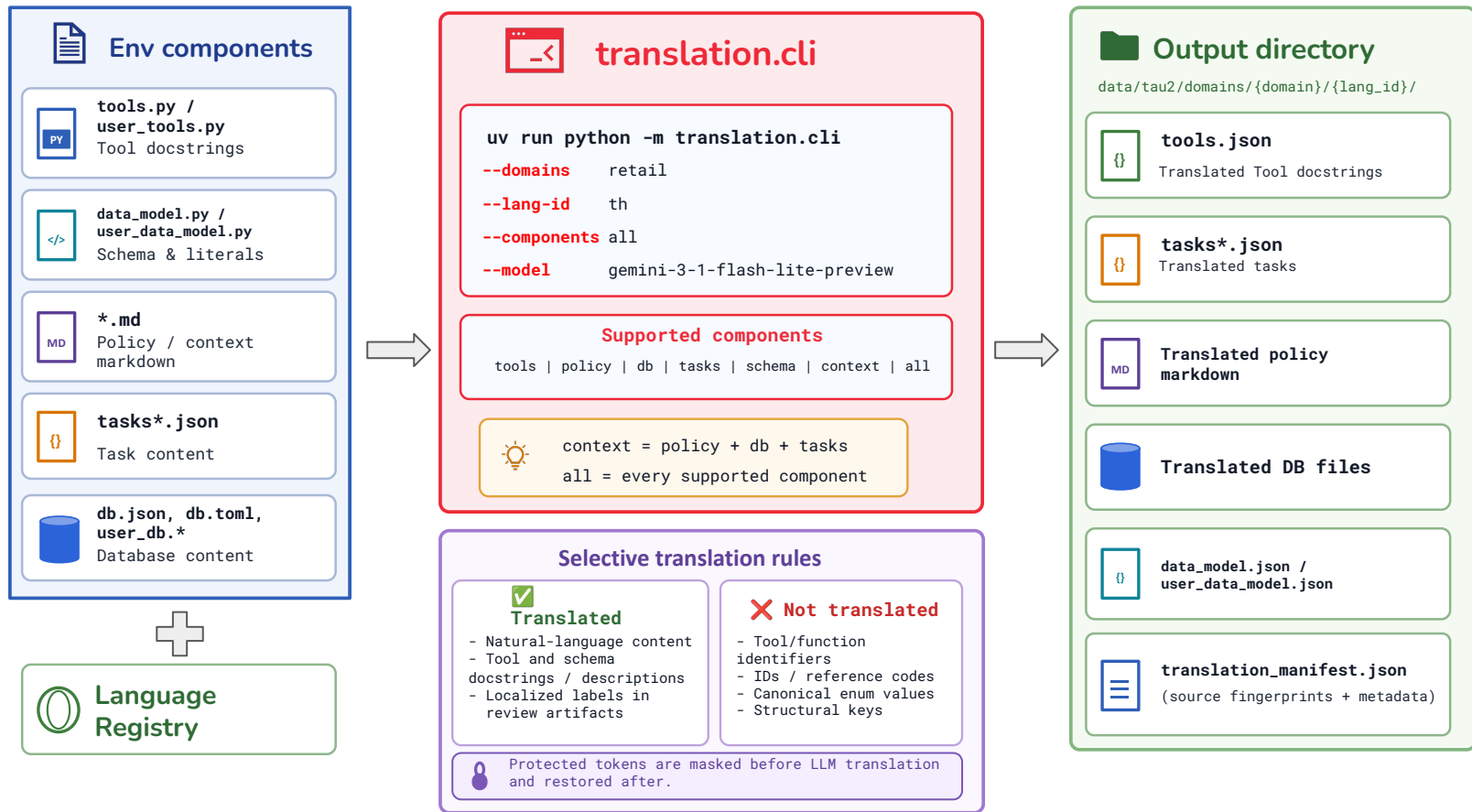
What we're curious about?

Research questions are centered around the agent adaptability in cross- and multilingual environments.



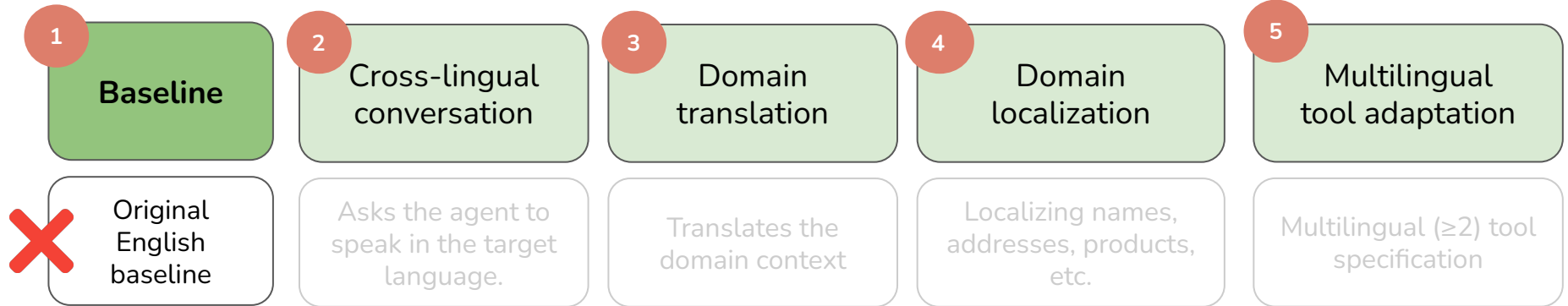
Simulate agents that interact with APIs documentation from regions with different languages

Translation pipeline (scenario 3+4+5)



Data curation: what to do in each scenario?

No translation needed

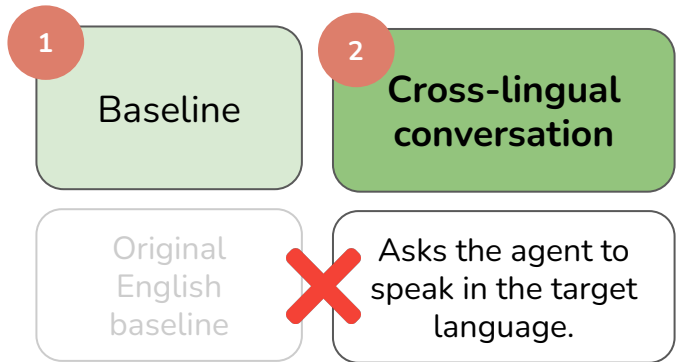


Run the original domain **without** modifying source artifacts, tools and situation guidelines (.md)

 Env components

Data curation: what to do in each scenario?

No translation needed



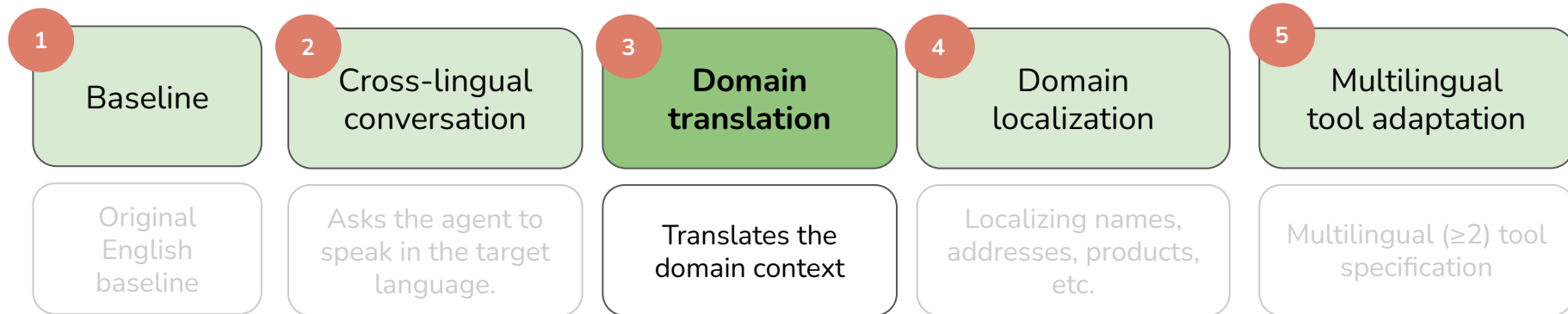
```
{
  "role": "assistant",
  "content": "{\n  \"sukses\": true,\n  \"pesan\": \"Alamat pengiriman untuk pesanan #W2702727 berhasil diperbarui sesuai permintaan Anda. Berikut ringkasan pembaruan.\",\n  \"pesanan_diperbarui\": {\n    \"order_id\": \"#W2702727\",\n    \"status\": \"pending\",\n    \"alamat_baru\": {\n      \"address1\": \"1234 Elm St\",\n      \"address2\": \"\",\n      \"city\": \"Springfield\",\n      \"state\": \"IL\",\n      \"zip\": \"62701\",\n      \"country\": \"USA\"\n    },\n    \"items\": [\n      {\n        \"name\": \"Bookshelf\",\n        \"item_id\": \"7373893106\",\n        \"price_usd\": 531.22\n      },\n      {\n        \"name\": \"Yoga Mat\",\n        \"item_id\": \"2733768059\",\n        \"price_usd\": 94.38\n      }\n    ],\n    \"total_amount_usd\": 625.6,\n    \"payment_method_id\": \"credit_card_3599838\",\n    \"catatan_penting\": [\n      \"Perubahan alamat berhasil dan tidak mengubah total pesanan. Total tetap $625.60 kecuali ada biaya tambahan yang muncul kemudian (tidak tercatat sekarang).\",\n      \"Status pesanan tetap 'pending'. Sistem tidak menjadwalkan pembatalan otomatis di masa depan. Jika Anda ingin membatalkan #W2702727 nanti (mis. jika masih pending setelah 5 hari), Anda harus menghubungi kami lagi dan memberikan konfirmasi pembatalan pada saat itu.\",\n      \"Jika Anda ingin, saya dapat membatalkan pesanan sekarang. Anda sebelumnya meminta pembatalan otomatis jika tidak terproses dalam 5 hari (2014 saya tidak dapat melakukan pembatalan otomatis; mohon beri instruksi eksplisit (jawab 'batalkan sekarang' untuk membatalkan sekarang).\",\n      \"langkah selanjutnya\": \"Beritahukan jika Anda ingin: (1) membatalkan #W2702727 sekarang, atau (2) membatalkan #W8268610 sekarang (jika ya, berikan alasan 'no longer needed' atau 'ordered by mistake'), atau (3) ada perubahan lain yang ingin dilakukan.\"\n    ]\n  }",
  "role": "user",
  "content": "Saya sudah menerima konfirmasi bahwa alamat untuk pesanan #W2702727 telah berhasil diubah ke 1234 Elm St, Springfield, IL, 62701. Terima kasih atas pembaruan tersebut.\n\nSaat ini, saya memilih untuk **tidak** membatalkan pesanan #W2702727 sekarang. Namun, saya ingin menegaskan kembali bahwa jika pesanan tersebut **masih dalam status pending setelah 5 hari dari sekarang**, maka saya akan menganggap bahwa layanan Anda tidak dapat memprosesnya tepat waktu, dan saya akan menghubungi kembali untuk meminta pembatalan serta pengembalian dana sebesar $625.60 ke kartu kredit yang digunakan.\n\nUntuk pesanan #W8268610, saya tidak ingin melakukan pembatalan atau perubahan apa pun saat ini.\n\nTidak ada permintaan tambahan untuk saat ini. Saya hanya ingin memastikan bahwa perubahan alamat sudah benar-benar tercatat dan tidak akan berdampak pada total biaya pesanan.\n\n##STOP###",
  "role": "assistant",
  "content": null,
  "tool_calls": [
    {
      "id": "call_Sb35myQyX0WycAnHyv21UEdx",
      "name": "modify_pending_order_address",
      "arguments": {
        "order_id": "#W2702727",
        "address1": "1234 Elm St",
        "address2": "",
        "city": "Springfield",
        "state": "IL",
        "country": "USA",
        "zip": "62701"
      }
    }
  ],
  "requestor": "assistant"
}
```

Add **prompts** to simulation guidelines (.md):

```
**Important: You must always respond in {language}. Do not use any other language.**  
**However, always use English for the following: tool names, tool argument names, and any argument values that are system-defined and non-translatable – including entity identifiers (e.g., IDs, reference codes, alphanumeric keys), enumerated constants (e.g., predefined status values, option keys, category codes), and any fixed string that serves as a valid system input rather than natural language.**
```

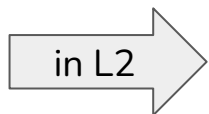
Data curation: what to do in each scenario?

Each component is translated using our translation pipeline.



We translate **user-agent domain context** into the target language while preserving entity consistency.

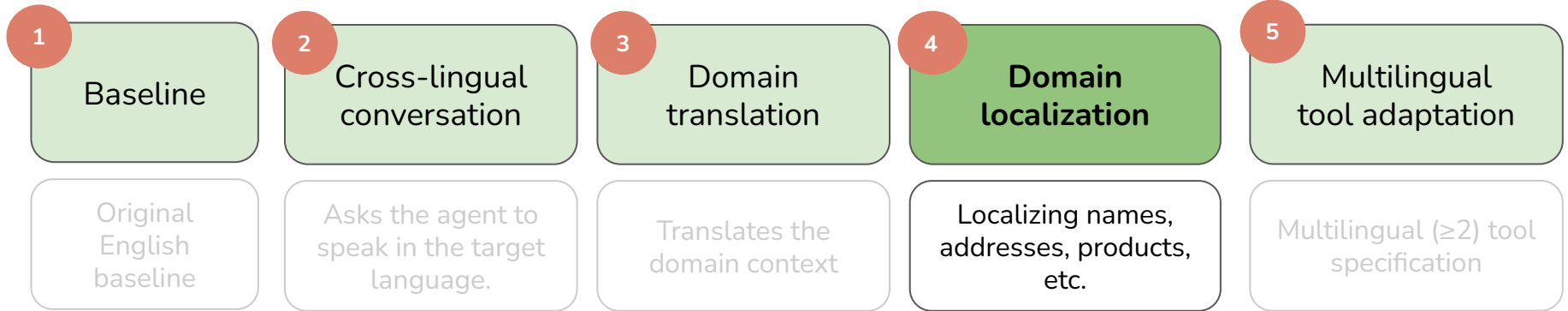
- Policy
- User profile
- Task description



→ Can the agent maintain policy and state reasoning when the relevant evidence is no longer in English?

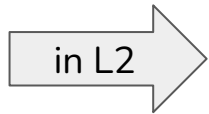
Data curation: what to do in each scenario?

Each component is translated using our translation pipeline.



We translate **user-agent domain context** into the target language while preserving entity consistency.

- Policy
- User profile
- Task description



Substitute **cultural specifics** (e.g., products, addresses) with plausible regional alternatives

→ Can the agent understand and maintain performance when the interaction artifacts are localized?

Data curation: what to do in each scenario?

Original Tools in English

```
@is_tool(ToolType.READ)
def get_product_details(self, product_id: str) ->
Product:
    """Get the inventory details of a product.

Args:
    product_id: The product id, such as
'6086499569'. Be careful the product id is
different from the item id.

Returns:
    Product: The product details.

Raises:
    ValueError: If the product is not found.
    """
    product = self._get_product(product_id)
    return product
```



eval

Tools in SEA Language

```
@is_tool(ToolType.READ)
def get_product_details(self, product_id: str) ->
Product:
    """รับรายละเอียดสต็อกของสินค้า.

Args:
    product_id: รหัสสินค้า เช่น '6086499569'. โปรดระวัง
ว่ารหัสสินค้าจะแตกต่างจากรหัสรายการ.

Returns:
    Product: รายละเอียดสินค้า.

Raises:
    ValueError: หากไม่พบสินค้า.
    """
    product = self._get_product(product_id)
    return product
```



translation
pipeline

eval

Tools in bi, tri, multi-lingual

5

**Multilingual
tool adaptation**

Multilingual (≥ 2) tool
specification



mixing
tools

We translate Tool descriptions while keeping the conversation and domain context in English
→ Can the model ground function use in multilingual documentation without changing the rest of the task?

Experimental setup



Model selections

User model:

Qwen3 235B A22B

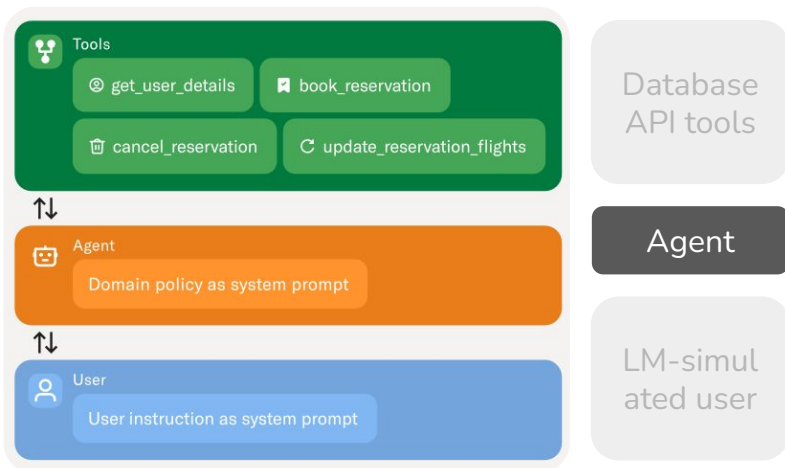
<https://artificialanalysis.ai/methodology/intelligence-benchmarking#tau2-bench-telecom>

 Artificial Analysis

τ^2 -Bench Telecom

- **Description:** Benchmark developed by Sierra for conversational AI agents in 'dual control' scenarios with language models simulating both agent and user roles to test planning, tool use, and guidance/communication
- **Paper:** <https://arxiv.org/abs/2506.07982>
- **Dataset:** <https://github.com/sierra-research/tau2-bench>
- **Implementation:**
 - The 'telecom' domain introduced in τ^2 -Bench contains 114 tasks (subsampling from a total 2,285 programmatically generated tasks), with varying 'intents' describing if the task is related to service, mobile data, or MMS issues. We evaluate the telecom domain in full with 3 repeats per task, and report the score using pass@1 scoring as the average of the 3 attempts
 - In this benchmark, the outcome 'world state' decides whether the agent succeeded - for example, whether the user's cell phone data is functioning after the agent completes the task
 - The full τ^2 -Bench suite includes 3 execution modes with varying planning and communication levels in ablation studies; we implement the 'default' dual control mode with fully simulated and separate user and assistant agents
 - We use [Qwen3 235B A22B 2507 \(Non-reasoning\)](#) for the user agent simulator to ensure consistent checkpoint availability and full control over inference settings alongside strong base intelligence
 - We apply a constraint on execution to limit steps to a maximum of 100 per task repeat

Experimental setup



Model selections

User model:

Qwen3 235B A22B

Agent model:

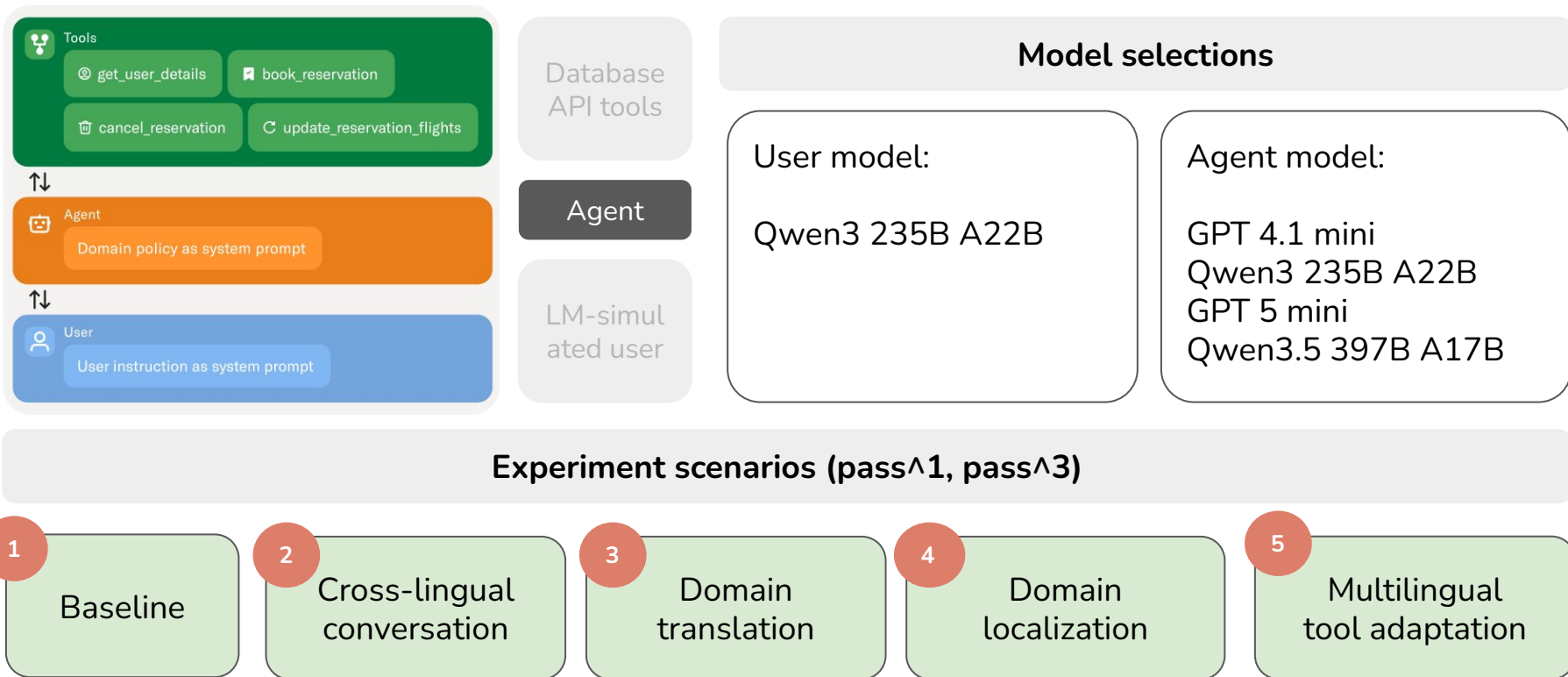
GPT 4.1 mini

Qwen3 235B A22B

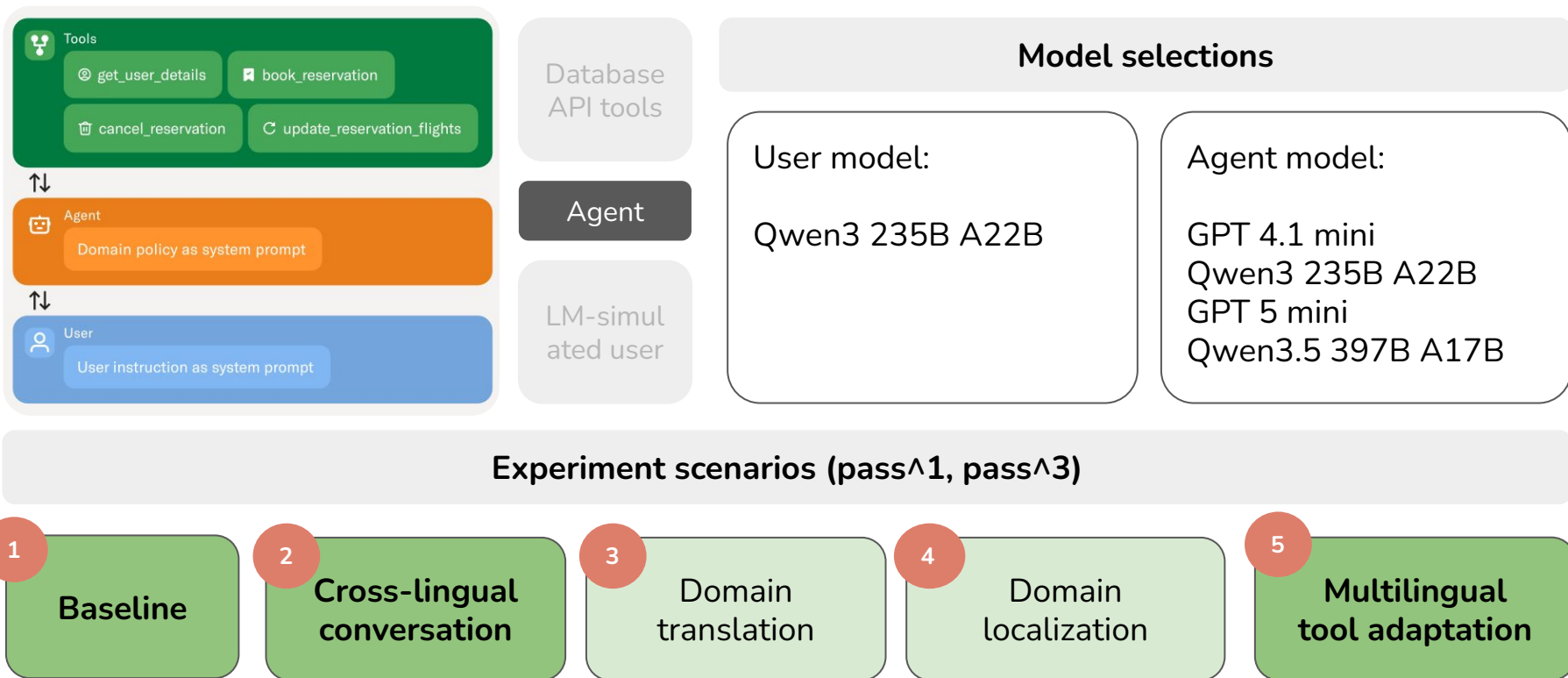
GPT 5 mini

Qwen3.5 397B A17B

Experimental setup



Experimental setup



Preliminary result: English Baseline

User model: Qwen3-235b-a22b-2507

Domain	Agent model	Pass^1	Pass^2	Pass^3
Airline	gpt-5.4-2026-03-05	0.607	0.493	0.440
	gpt-4.1-mini	0.380	x	x
Telecom	gpt-5.4-2026-03-05	0.459	0.310	0.237
	gpt-5-mini	0.664	0.532	0.474
	gpt-4.1-mini	0.518	x	x

Domain	Agent model	Pass^1	Pass^2	Pass^3
Retail	gpt-5.4-2026-03-05	0.667	0.529	0.447
	gpt-4.1-mini	0.509	x	x
Banking	gpt-5.4-2026-03-05	0.127	0.069	0.052
	gpt-5-mini	0.124	0.065	0.041
	qwen3-235b-a22b-2507	0.069	0.041	0.031

Key takeaways

- GPT family is arch-superior in performance compared with Qwen3-235b-a22b-2507 in Banking Domain
- GPT-5 mini is a valuable model to use as an agent.

Preliminary result: English Baseline

User model: Qwen3-235b-a22b-2507

Domain	Agent model	Pass^1	Pass^2	Pass^3
Airline	gpt-5.4-2026-03-05	0.587	0.473	0.420
	gpt-4.1-mini	0.380	x	x
Telecom	gpt-5.4-2026-03-05	0.459	0.310	0.237
	gpt-5-mini	0.664	0.532	0.474
	gpt-4.1-mini	0.518	x	x

Domain	Agent model	Pass^1	Pass^2	Pass^3
Retail	gpt-5.4-2026-03-05	0.667	0.529	0.447
	gpt-4.1-mini	0.509	x	x
Banking	gpt-5.4-2026-03-05	0.124	0.069	0.052
	gpt-5-mini	0.124	0.065	0.041
	qwen3-235b-a22b-2507	0.069	0.041	0.031

Key takeaways

- GPT family is arch-superior in performance compared with Qwen3-235b-a22b-2507 in Banking Domain
- GPT-5 mini is a valuable model to use as an agent.

Preliminary result: Cross-lingual conversation

Results

User: Qwen3-235b-a22b-2507

Agent: GPT-5-mini

Domain	Language				
	Vietnamese	Thai	Indonesian	Chinese	Filipino
airline	Pass^1=0.640 Pass^2=0.500 Pass^3=0.420	Pass^1=0.593 Pass^2=0.473 Pass^3=0.400	Pass^1=0.627 Pass^2=0.447 Pass^3=0.360	Pass^1=0.620 Pass^2=0.460 Pass^3=0.380	Pass^1=0.620 Pass^2=0.527 Pass^3=0.480
telecom	Pass^1=0.561 Pass^2=0.430 Pass^3=0.368	Pass^1=0.605 Pass^2=0.436 Pass^3=0.342	To Be Done		
retail	Pass^1=0.696 Pass^2=0.570 Pass^3=0.500	Pass^1=0.699 Pass^2=0.538 Pass^3=0.430	Pass^1=0.684 Pass^2=0.547 Pass^3=0.456	Pass^1=0.675 Pass^2=0.506 Pass^3=0.412	Pass^1=0.655 Pass^2=0.518 Pass^3=0.430
banking	To Be Done				

Key takeaways

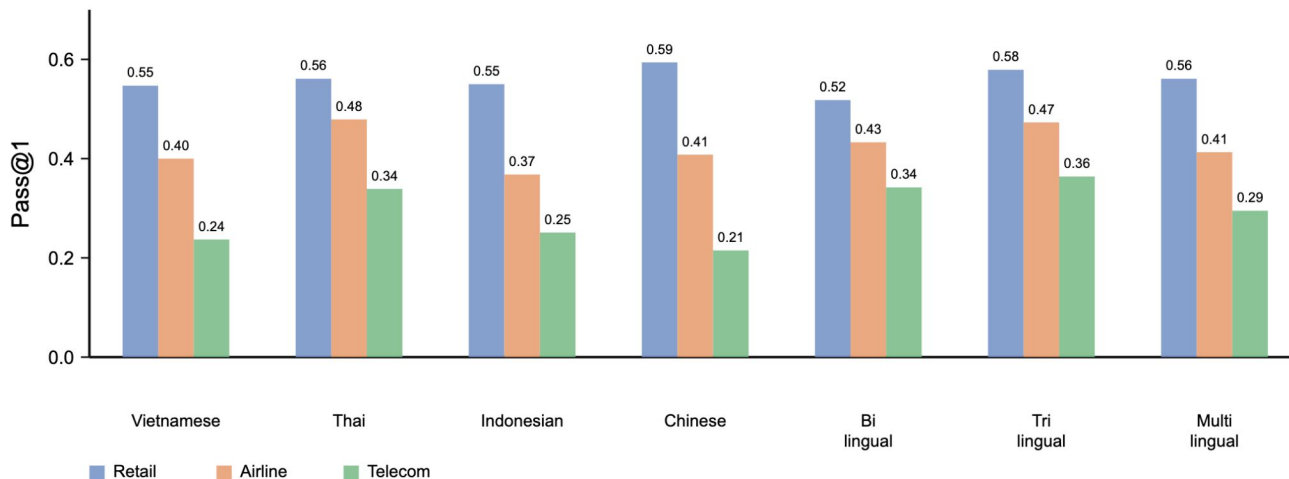
- **Strong initial success:** Cross-lingual performance on the first trial (Pass^1) are more or less same as Scenario 1.
- **Multi-trial degradation:** Performance drops steeply in subsequent trials (Pass^2 and Pass^3).
- **English stability:** The LLM maintains consistent performance across repeated attempts better in English.

Preliminary result: Multiling tool adaptation (Pass@1)

Results

User Model:
qwen3-235b-a22b
-instruct-2507

Agent Model:
qwen3-235b-a22b
-instruct-2507



Key takeaway:

- Retail consistently leads (0.52–0.59), followed by Airline (0.37–0.48), while Telecom is lowest (0.21–0.36).
- Performance is relatively stable across languages, with slight gains in tri-lingual configurations and a small drop in bi-lingual settings.

Preliminary result: Multilingual tool adaptation

Results

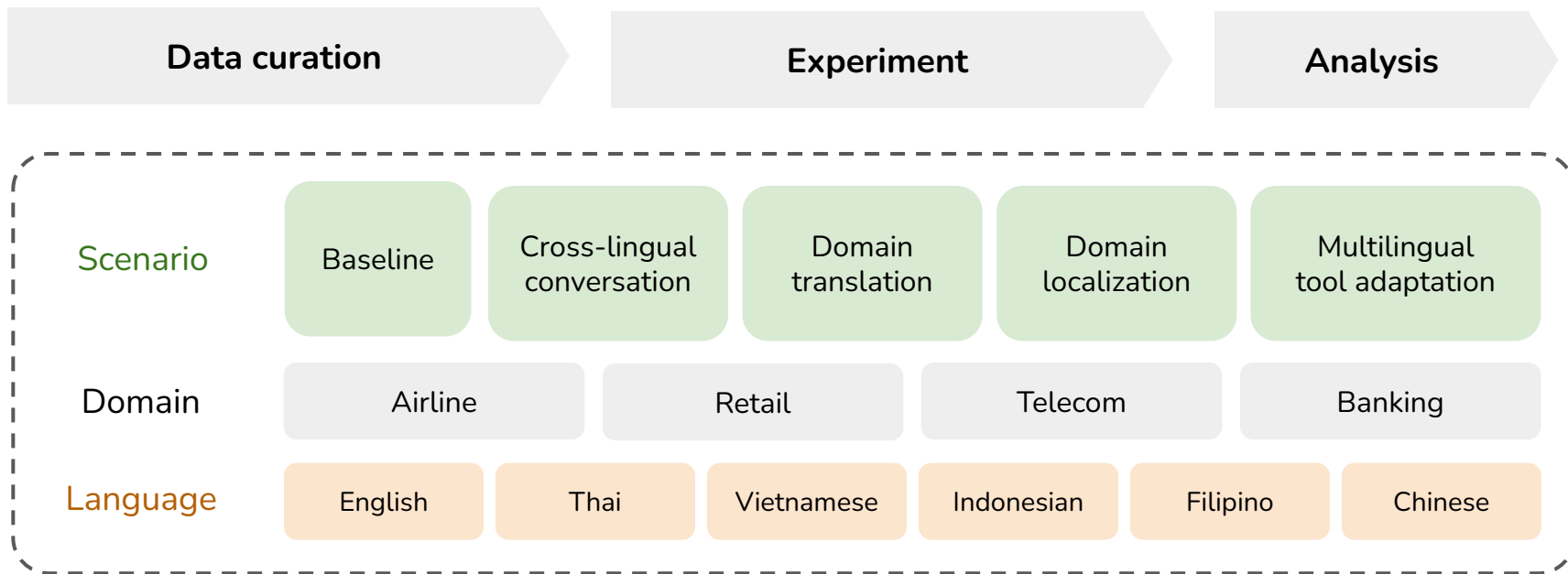
User Model:
qwen3-235b
-a22b-instruc
t-2507
Agent Model:
qwen3-235b
-a22b-instruc
t-2507

Domain	Vietnamese	Thai	Indonesian	Chinese	Bi- lingual	Tri- lingual	Multi- lingual
Retail	pass@1 0.547	pass@1 0.561	pass@1 0.550	pass@1 0.594	pass@1 0.518	pass@1 0.579	pass@1 0.561
	pass@2 0.412	pass@2 0.421	pass@2 0.418	pass@2 0.433	pass@2 0.357	pass@2 0.418	pass@2 0.389
	pass@3 0.351	pass@3 0.351	pass@3 0.351	pass@3 0.360	pass@3 0.272	pass@3 0.342	pass@3 0.298
Airline	pass@1 0.400	pass@1 0.479	pass@1 0.368	pass@1 0.408	pass@1 0.433	pass@1 0.473	pass@1 0.413
	pass@2 0.300	pass@2 0.382	pass@2 0.236	pass@2	pass@2 0.300	pass@2 0.340	pass@2 0.240
	pass@3	pass@3	pass@3 0.167	pass@3	pass@3 0.260	pass@3 0.300	pass@3 0.160
Telecom	pass@1 0.237	pass@1 0.339	pass@1 0.251	pass@1 0.215	pass@1 0.342	pass@1 0.364	pass@1 0.295
	pass@2	pass@2 0.208	pass@2	pass@2	pass@2 0.181	pass@2 0.199	pass@2 0.190
	pass@3	pass@3 0.140	pass@3	pass@3	pass@3 0.114		

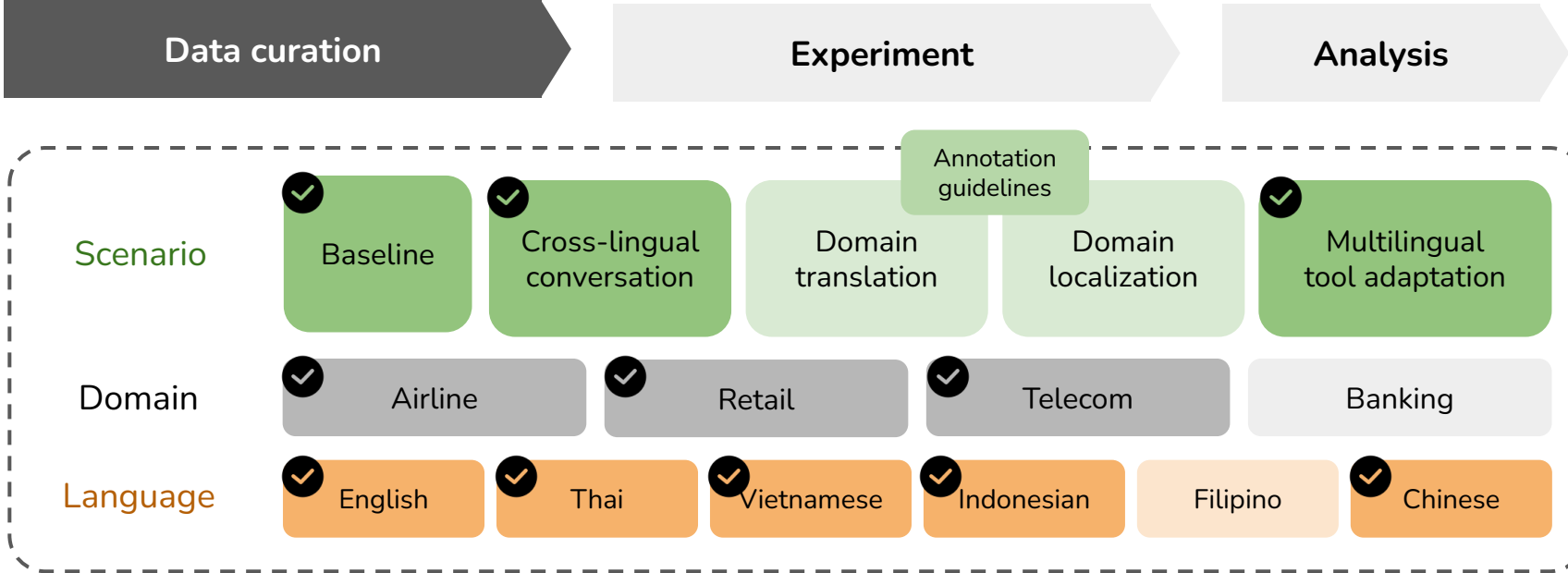
Key takeaway

Performance across domains shows Chinese leading overall, Thai strongest in Airline, Retail as the easiest domain, Telecom as the most challenging, and diminishing returns when scaling beyond tri-lingual setups.

What's next?



What's next?



What's next?

