ICML 2026

Seoul, South Korea

EffGen: Enabling Small Language Models as Capable Autonomous Agents

Gaurav Srivastava[1], Aafiya Hussain[1], Chi Wang[2], Yingyan (Celine) Lin[3], Xuan Wang[1]
[1] Department of Computer Science, Virginia Tech, USA
[2] Google DeepMind, USA
[3] Department of Computer Science, Georgia Tech, USA

VIRGINIA TECH

DeepMind

The Problem

What is the current gap?

- Existing agentic systems are built for large models, not resource-constrained small models

Most frameworks target:
- GPT
- Claude
- Gemini
- And other popular LLM providers

But

LLM are:
- expensive
- Hard to deploy locally
- privacy-sensitive

The Problem

Why current agent frameworks fall short for SLMs?

Small Language Models have:

- have limited context windows
- weaker reasoning
- poorer instruction following

But existing frameworks do not optimize for these limitations!

Can we build an agent framework specifically designed for SLM constraints?

Why EFFGEN?

EFFGEN as an SLM-First Agent Framework

Redesigning the framework around small-model limitations instead of adapting LLM systems

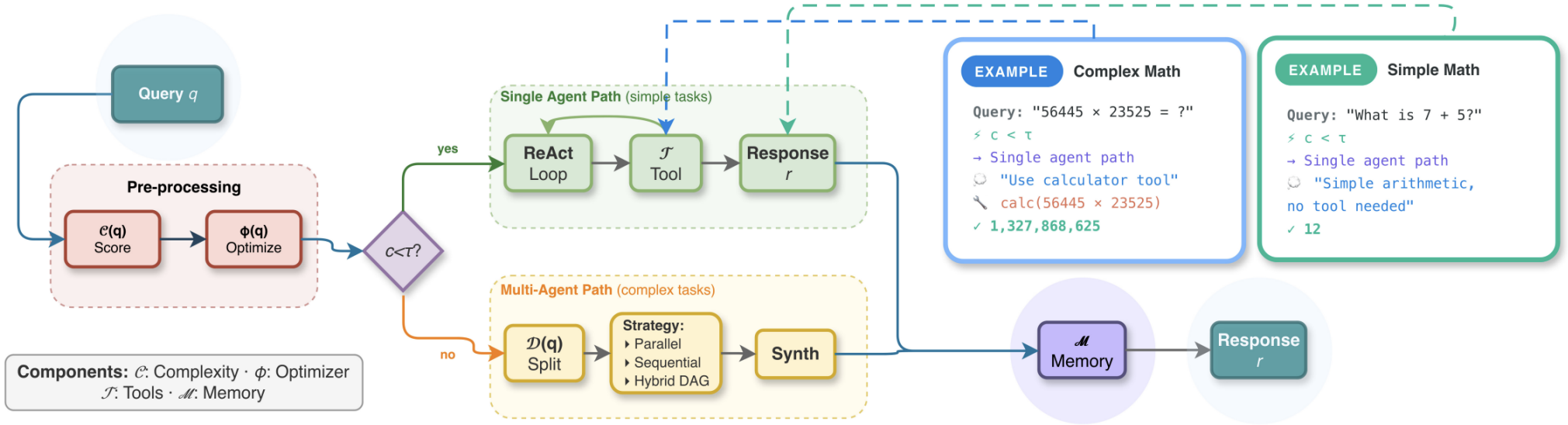
Main Contributions:

- Prompt Optimization
- Complexity-Based Routing
- Intelligent Task Decomposition
- Three-Tier Memory System
- Multi-Protocol Communication

Core Philosophy: SLM constraints are treated first-class design requirements

Overall Framework

EFFGEN Execution Pipeline



Simple Task – Handled Directly, Complex Task - Decomposed

Prompt Optimization

Making Prompts Easier for Small Models

- SLMs struggle with long prompts, verbose instructions, complex formatting

SLMS struggle with:
- long prompts
- verbose instructions
- complex formatting

EFFGEN Optimizations:
- compression
- simplification
- redundancy removal
- bullet formatting
- context truncation

We achieved **70–80%** prompt compression with **8–11%** performance improvement

Complexity Routing

Adaptive Execution for Different Task Difficulties

- Uses task complexity and memory-aware execution to improve efficiency

Complexity analyzed based on:

- Task Length
- Requirements
- Domain Breadth
- Tool Needs
- Reasoning Depth

Routing Decisions:

- Simple → single-agent
- Complex → decomposition
- Very complex → hierarchical execution

- Routing improves performance by **6–11% over always-single or always-multi strategies**
- Reduces wasted computation by **23%**

Three-tier Memory

Combining short-term context, long-term storage, and semantic retrieval for SLMs

- SLMs struggle with limited context windows and long-horizon reasoning.
- EffGen solves this using:
  1. Short-Term memory: maintains recent conversational context
  2. Long-Term memory: stores past events and interactions
  3. Vector memory: Semantic retrieval of relevant information

- Improves memory benchmark performance by **5-15%**
- Retrieves **6 relevant** segments from 15K-token conversations in **23ms**

Results

Testing across Qwen models of different scales

Model	Framework	Calculator			Math Reasoning (coding tools)			Agentic Benchmarks		Memory		Retrieval			Avg
		GSM8K	GSM-PLUS	MATH-500	BB-Easy	BB-Med	BB-Hard	GAIA	SimpleQA	LoCoMo	LongMemEval	ARC-C	ARC-E	CSQA	
Qwen2.5 (1.5B)	Raw Model	68.01	42.83	28.60	35.98	9.60	2.86	3.12	4.00	5.13	22.88	67.24	82.70	72.73	34.28
	LangChain	66.86	37.83	23.20	49.39	21.60	1.43	3.12	12.00	8.56	15.34	59.13	73.78	54.30	32.81
	AutoGen	67.85	37.62	26.19	52.95	28.40	1.67	6.25	10.00	9.22	18.73	53.75	65.99	57.82	34.33
	Smolagents	50.41	26.08	21.25	56.39	30.40	1.43	3.12	18.00	8.42	22.15	53.85	37.63	32.41	27.81
	EFFGEN	**71.63**	**50.25**	**36.00**	**73.66**	**38.40**	**7.92**	**9.38**	**40.00**	**14.19**	**30.73**	**78.34**	**91.65**	**74.62**	**47.44**
Qwen2.5 (3B)	Raw Model	82.64	62.04	55.00	42.05	18.20	7.33	6.25	2.00	17.40	29.25	79.28	90.60	75.02	43.62
	LangChain	82.83	58.33	48.80	52.68	29.20	5.71	9.38	12.00	13.42	11.62	74.49	83.88	70.84	42.55
	AutoGen	79.30	60.21	51.80	58.78	26.80	7.14	6.25	14.00	11.81	20.57	73.04	85.14	72.65	43.65
	Smolagents	69.07	49.62	10.91	63.41	28.80	8.57	9.38	22.00	7.89	19.54	36.69	43.43	41.03	31.56
	EFFGEN	**84.83**	**63.62**	**59.20**	**81.71**	**35.60**	**21.67**	**15.62**	**58.00**	**21.58**	**32.29**	**86.10**	**93.21**	**85.02**	**56.80**
Qwen2.5 (7B)	Raw Model	90.75	**72.54**	65.80	59.68	27.60	15.42	12.50	6.00	22.35	29.88	83.79	90.91	82.80	50.77
	LangChain	84.38	67.79	41.20	72.20	45.60	12.86	9.38	34.00	9.02	16.32	86.95	90.99	79.12	49.99
	AutoGen	90.30	71.25	64.60	77.80	43.20	7.14	12.50	46.00	16.44	27.83	76.28	80.39	58.64	51.72
	Smolagents	84.00	64.58	16.20	80.24	49.20	24.17	9.38	32.00	10.25	27.47	71.25	81.36	76.33	48.19
	EFFGEN	**91.28**	68.12	**66.80**	**89.02**	**54.80**	**28.57**	**21.87**	**76.00**	**25.32**	**35.34**	**87.88**	**91.92**	**83.05**	**63.07**
Qwen2.5 (14B)	Raw Model	93.78	75.17	67.80	64.63	38.80	24.17	12.50	8.00	26.92	30.94	88.99	92.97	83.37	54.46
	LangChain	89.99	70.92	53.20	74.71	47.20	29.58	12.50	46.00	19.00	17.68	**91.13**	93.90	81.49	55.95
	AutoGen	94.09	74.83	66.00	77.07	51.20	24.29	15.62	58.00	17.15	27.72	89.42	92.89	79.69	59.07
	Smolagents	90.45	71.17	20.43	86.27	48.80	34.29	9.38	46.00	13.04	24.34	90.87	93.14	82.80	54.69
	EFFGEN	**94.84**	**76.62**	**69.60**	**92.27**	**57.60**	**46.37**	**18.75**	**72.00**	**27.94**	**36.64**	89.86	**94.39**	**86.00**	**66.38**
Qwen2.5 (32B)	Raw Model	95.60	77.42	73.00	71.56	48.40	26.53	15.62	14.00	**31.91**	30.12	92.92	94.32	**86.81**	58.32
	LangChain	95.07	75.83	59.20	88.05	60.00	27.14	18.75	54.00	25.07	18.66	**93.34**	93.94	84.28	61.03
	AutoGen	94.54	76.96	72.80	92.44	63.60	30.67	21.87	70.00	30.37	28.22	89.25	92.51	81.65	64.99
	Smolagents	93.40	74.83	68.00	94.63	**68.80**	45.83	15.62	52.00	17.07	27.22	93.26	94.44	84.93	63.85
	EFFGEN	**95.75**	**78.38**	**75.40**	**96.67**	64.20	**58.86**	**28.12**	**84.00**	31.04	**35.64**	92.50	**95.29**	86.80	**70.97**

EffGen outperforms AutoGen, LangChain and SmolAgents

Ablations

Component contribution analysis across Qwen model scales

Configuration	1.5B Acc	1.5B Δ	7B Acc	7B Δ	32B Acc	32B Δ
Full EFFGEN	47.44	–	63.07	–	70.97	–
– Prompt Optim.	36.21	−11.2	54.18	−8.9	68.54	−2.4
– Complexity Routing	43.87	−3.6	56.84	−6.2	63.12	−7.9
– Task Decomp.	44.12	−3.3	58.42	−4.7	65.48	−5.5
– Memory System	45.68	−1.8	59.71	−3.4	67.23	−3.7
– All (Raw ReAct)	34.28	−13.2	50.77	−12.3	58.32	−12.7

- Smaller models benefit the most from optimized framework design.
- Complexity routing is more useful for large models.

Acknowledgements and Resources

Supported by

- NSF, NAIRR Pilot (PSC Neocortex, NCSA Delta), Cisco Research, NVIDIA, Amazon, Commonwealth Cyber Initiative, Amazon-VT Center for ML, Sanghani Center, VT Innovation Campus

Resources

- Code: https://github.com/ctrl-gaurav/effGen
- Package: `pip install effgen`
- Website/Documentation: https://effgen.org/

Contact

- gks@vt.edu | xuanw@vt.edu

Better coordination may be more important than simply scaling model size.

Thanks to

Wang's Lab @ VT & Google

Deepmind & GaTech University

↓ Paper Link ↓

↓ GitHub ↓

Author Details

Email: gks@vt.edu

↓ Personal Page ↓

↓ LinkedIn ↓

ICML 2026

Seoul, South Korea

Thank You!