TIGRIS: Integrating Policy Knowledge Graphs and Theory of Change for Institutionally Grounded Policy Impact Simulation

Haein Yeo[1,*], **Junghyun Kim**[2,*], **Beejin Son**[1], **Kyungsik Han**[1,2,†]

[1] Department of Artificial Intelligence, Hanyang University
[2] Department of Data Science, Hanyang University
Seoul, Republic of Korea
{haeinyeo,junghyuning,beejinson,kyungsikhan}@hanyang.ac.kr
[*]Equal contribution. [†]Corresponding author.

Abstract

Anticipating the social impacts of a policy prior to implementation requires *institutional consistency*, ensuring that multi-agent reasoning should be aligned with the rules specified in policy documents. However, existing LLM-based multi-agent simulations struggle to maintain such fine-grained institutional rules when policy documents are provided only as prompt context. In this paper, we propose **TIGRIS**, a **T**heory-guided framework for **I**nstitutionally **GR**ounded policy **I**mpact **S**imulation. TIGRIS operates by (1) constructing a structured policy knowledge graph from natural-language policy documents using the Institutional Analysis and Development framework to define institutional positions and Institutional Grammar 2.0 to encode behavioral rules, and (2) instantiating actor nodes as persona agents while anchoring them to a Theory of Change framework to map the causal pathways of policy impacts. Across four AI talent development policies and three LLMs, TIGRIS achieves the lowest macro-average MAPE. Additional experiments on public-health and welfare policies assess cross-domain transfer, while identifier masking evaluates sensitivity to policy identifiers.

1 Introduction

Anticipating the social impacts of a policy prior to implementation has long been a central challenge for experts in public policy and policy informatics (Johnston, 2015; Dawes and Janssen, 2013). The impacts of a policy emerge dynamically as multiple stakeholders adapt their choices and interactions under the specific institutional rules—such as roles, authority, obligations, and exceptions—specified in policy documents (Ostrom, 2005; McGinnis, 2011). Therefore, analyzing the impact of policies before implementation fundamentally requires *institutional consistency*, ensuring that the modeling process and result interpretation should remain aligned with the institutional rules described in policy documents.

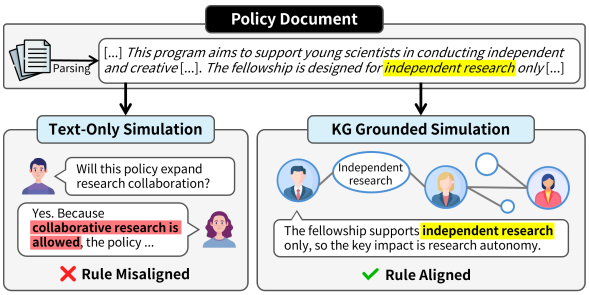


Figure 1: Motivation of TIGRIS. Without an explicit institutional structure, LLM agents may contradict policy rules (left), whereas a policy knowledge graph keeps their reasoning aligned with these rules (right).

Recent advances in large language models (LLMs) have led to their increasing deployment in multi-agent simulations for exploring stakeholder responses and social interactions during policy implementation, empowered by their reasoning and role-playing capabilities (Park et al., 2023; Li et al., 2024; Mou et al., 2024b). Achieving such *institutional consistency* requires operationalizing the fine-grained institutional rules and constraints within policy documents into an explicit, structured representation that anchors agents' reasoning processes throughout a simulation (Liu et al., 2024; Hsieh et al., 2024) (Figure 1).

In this context, previous studies have explored various methodologies for executing policy simulations or structuring policy-relevant knowledge; however, each also introduces limitations. First, LLM-based social simulation enables researchers to explore societal responses by generating open-ended discussions and interactions among agents (Park et al., 2023; Yang et al., 2024; Zhou et al., 2025). However, policy documents are often used as baseline scenario descriptions, making it difficult to ensure that agents adhere to the roles, authority, obligations, and exceptions specified in

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33008–33031
October 24-29, 2026 ©2026 Association for Computational Linguistics

those documents throughout a simulation (Mou et al., 2024b; Mu et al., 2024). Second, knowledge graph construction has proven effective for structuring actors and relations in policy documents (Mou et al., 2024a; Li et al., 2025; Zhang and Soh, 2024). Nevertheless, conventional entity-relation representations struggle to capture the normative and conditional nature of the policy domain, where institutional positions, authority, conditions, obligations, and exceptions operate together. Third, Generative Agent-Based Modeling (GABM) can model how collective macro-level policy impacts emerge by simulating micro-level agent behaviors based on agents, environments, interactions, and rules (Vezhnevets et al., 2023). However, most existing studies rely on rule sets specified in advance by researchers, failing to address how complex, natural-language policy documents can be transformed into structured representations for institutionally consistent and causally grounded simulation.

Addressing these barriers requires systematically integrating theoretical perspectives from the policy domain into the reasoning process of LLM-based simulations, so that established theoretical criteria—rather than unconstrained prompt-based reasoning—guide what models reference, how they judge, and how they verify their reasoning. Recent work has begun to explore such theory-guided reasoning, including principle-based alignment methods (Bai et al., 2022; Madaan et al., 2023) and frameworks that incorporate behavioral theory into user modeling (Noh et al., 2025, 2026). While LLM-based simulation is increasingly recognized as a viable tool for exploring stakeholder dynamics and social impacts of policy (Li et al., 2024; Mou et al., 2024b), a systematic methodology for translating the institutional characteristics of policy documents into such simulations remains underexplored.

To fill this gap, in this paper, we present **TIGRIS**, an end-to-end **T**heory-guided framework for **I**nstitutionally **GR**ounded policy **I**mpact **S**imulation. Starting from natural-language policy documents, TIGRIS integrates three foundations from policy theory into multi-agent simulation: *Institutional Analysis and Development* (IAD) (Ostrom, 2011) for institutional interaction structures, *Institutional Grammar 2.0* (IG 2.0) (Frantz and Siddiki, 2021) for behavioral rules, and *Theory of Change* (ToC) for tracing causal policy pathways (Weiss et al., 1995). In the first stage, we construct a policy knowledge graph from natural-language policy documents; here, IAD defines the node structure by identifying key actors and institutional positions, while IG 2.0 establishes the relation structure by extracting rule components such as authority, obligation, condition, and exception. In the second stage, we instantiate these actor nodes as policy-grounded persona agents and organize their reasoning along a ToC result chain. As a result, TIGRIS uses institutional rules as a persistent, structured reference throughout the simulation, yielding traceable, theory-consistent pathways for estimating policy impacts.

We evaluate TIGRIS on four AI talent development policies using DeepSeek-V4-Flash (DeepSeek-AI, 2026), GPT-4o-mini (OpenAI, 2024), and Gemini-2.0-Flash (Google DeepMind, 2025). We further evaluate its cross-domain applicability on public-health and welfare policies using DeepSeek-V4-Flash and assess sensitivity to explicit policy identifiers by applying identifier masking to all methods on one representative policy from each of the three domains. TIGRIS achieves the lowest macro-average MAPE for each backbone in the primary evaluation. Our contributions are as follows:

- We develop TIGRIS, an end-to-end pipeline that operationalizes foundational policy theories within multi-agent-based simulations for reliable domain-aligned reasoning and decision-making.
- Through quantitative experiments spanning AI talent development, public health, and welfare, together with identifier masking, KG extraction validation, and qualitative analysis, we evaluate prediction accuracy, cross-domain transfer, and institutional grounding.

2 Related Work

2.1 LLM-Based Social Simulation for Policy Analysis

Policy simulation and agent-based modeling have long been used to examine how macro-level policy effects emerge from the micro-level interactions of heterogeneous actors under predefined environments and rules. Recent advances in LLM-based social simulation extend this line of work by using language-based agents to generate individual behaviors, social interactions, and emergent patterns in simulated environments (Yang et al., 2024;

Mou et al., 2024b). For example, Park et al. (2023) showed that LLM agents can produce long-term behavioral patterns and social interactions through architecture leveraging memory, reflection, and planning. Following this work, researchers have extended LLM-based social simulations to diverse domains, including online interactions, opinion formation, economic decision-making, and public resource allocation (Zhou et al., 2025; Chuang et al., 2024; Li et al., 2024). Generative Agent-Based Modeling (GABM) further integrates LLM-based decision-making into agent-based modeling components, systematically embedding actors, environments, interactions, rules, and time dynamics (Vezhnevets et al., 2023). Recent studies have applied GABM to evaluate economic policies, resource allocation, and large-scale social interventions (Ji et al., 2024; Mou et al., 2024b).

These advances make LLM-based simulation effective for exploring heterogeneous stakeholder responses to policy interventions. However, previous LLM-based social simulations often treat a policy merely as a scenario condition to which agents respond, rather than as a set of institutional rules that should guide their reasoning (Yang et al., 2024; Zhou et al., 2025). Previous GABM studies also tend to rely on environmental constraints and behavioral rule sets that are predefined by researchers (Vezhnevets et al., 2023; Li et al., 2024). As a result, most previous studies leave underexplored how to transform natural-language policy documents into structured representations that align agents' reasoning and interactions with institutional boundaries throughout a simulation. We address this gap by converting document-specified institutional rules into an explicit policy knowledge graph, establishing it as a persistent reference structure for multi-agent reasoning.

2.2 Knowledge-Structured Reasoning in Institutional Domains

Structuring external knowledge has emerged as a viable paradigm to ground LLM reasoning within policy and political domains (Pan et al., 2024). For example, frameworks, such as PEG (Mou et al., 2024a) and PAA (Li et al., 2025), show that political-domain knowledge can be externalized into structured representations and incorporated into LLM inference, rather than relying solely on the model's knowledge. While these studies motivate our use of structured knowledge, their grounding architectures primarily focus on political facts, actors, and agent profiles, rather than the institutional rules that govern policy implementation.

Policy documents specify how institutional actors should judge and act within defined roles and administrative relations; these rules should remain aligned with agents' reasoning and cross-agent interactions throughout the simulation. Conventional entity-relation knowledge graphs struggle to represent whether an actor is obligated, permitted, or prohibited to act under conditions or-else consequences (Zhang and Soh, 2024). Our work addresses this representation bottleneck by constructing a theory-grounded policy knowledge graph, combining IAD to capture institutional interaction structures and IG 2.0 to encode behavioral rules.

2.3 Theory-Guided Causal Reasoning in Multi-Agent Systems

Recent studies have introduced explicit structures to guide LLM agents' reasoning beyond unconstrained natural-language generation. Chai et al. (2025) improved multi-agent collaborative planning by combining a task graph with causal relations between sub-tasks. Hahm et al. (2025) deployed a causal influence diagram to guide agents to consider potential outcomes and risks during decision-making. The need for such structures is further underscored by LLMs' known limitations in causal reasoning. For example, Joshi et al. (2024) revealed that LLMs often incorrectly infer causal relations based on temporal order and positional cues, suggesting that policy impact analysis should disentangle temporal progression from true causal explanation. However, these studies do not sufficiently explain how institutional rules derived from policy documents are reflected across implementation stages and shape policy impacts.

To structure staged reasoning about policy impacts, we adopt Theory of Change (ToC), a theory-based evaluation approach that links interventions to expected societal outcomes through explicit assumptions (Weiss et al., 1995; Funnell and Rogers, 2011). In policy and program evaluation, ToC is conventionally operationalized as a sequential result chain spanning inputs, activities, outputs, outcomes, and impacts (OECD, 2022). We leverage this theoretical structure as a causal-temporal reasoning scaffold for our multi-agent simulation, organizing agents' judgements and predictions across progressive implementation stages while remaining grounded in document-derived institutional rules.

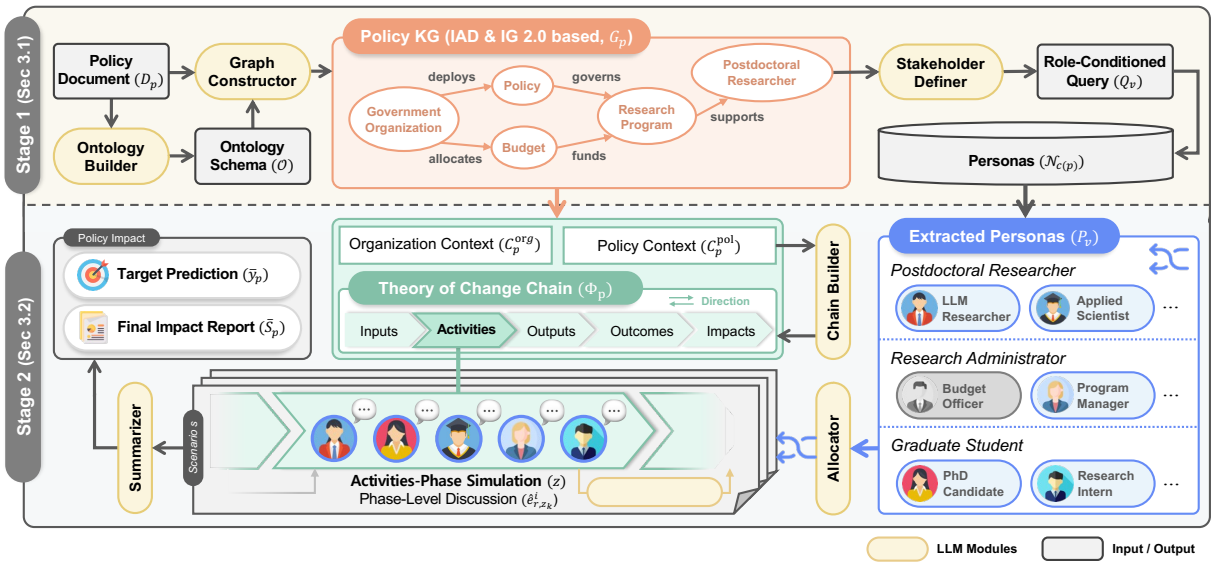


Figure 2: Overall framework of TIGRIS. Stage 1 constructs a policy knowledge graph from a policy document using IAD-based nodes and IG 2.0-based edges; Stage 2 instantiates persona-grounded agents and runs forward/backward ToC simulations (Activities phase shown for illustration).

3 Method

In this section, we introduce TIGRIS, a two-stage framework engineered to estimate policy impact indicators from policy documents (Figure 2): (1) Institutional Policy Knowledge Graph Construction (Sec 3.1) and (2) ToC-Guided Policy Impact Simulation (Sec 3.2). Pseudocode and prompt templates are provided in Appendices A and B.

3.1 Institutional Policy Knowledge Graph Construction

TIGRIS converts a policy document D_p into a policy knowledge graph G_p under an ontology schema:

$$\mathcal{O} = (\mathcal{T}_{\text{IAD}}, \mathcal{T}_{\text{IG}}, \mathcal{A}), \tag{1}$$

where $\mathcal{T}_{\text{IAD}}$ is the set of node types derived from the IAD framework, $\mathcal{T}_{\text{IG}}$ is the set of edge types derived from IG 2.0, and $\mathcal{A}$ defines attributes (e.g., eligibility criteria and activation conditions) that capture entity-level properties.

3.1.1 Policy Ontology

The IAD framework (Ostrom, 2011) models institutional interactions through action-situation components such as actors, positions, actions, information, control, costs and benefits, and outcomes. For graph construction, TIGRIS adopts five IAD-based node types: *Participant* (individual or organizational actors), *Position* (institutional roles assigned to participants), *Action* (policy-defined behaviors), *Information* (knowledge accessible to participants), and *Costs and Benefits* (resources, expenditures, or rewards). Among these, Participant and Position nodes serve as actor nodes used for agent construction.

IG 2.0 (Frantz and Siddiki, 2021) decomposes institutional statements using the ABDICO syntax, which extends the original ADICO grammar of Crawford and Ostrom (1995) with an explicit object component. In TIGRIS, each rule-grounded edge stores the corresponding ABDICO components as a rule statement, including the actor to whom the rule applies, the deontic modality, the aim or action, the object, activation conditions, execution constraints, and or-else consequences. The full edge type vocabulary and an illustrative ABDICO decomposition are provided in the Appendix F.

3.1.2 LLM-Based Graph Extraction

The input document D_p is an implementation plan released before policy execution, ensuring that post-implementation performance indicators are used exclusively for model evaluation. TIGRIS extracts G_p from D_p under the ontology constraint:

$$G_p = \text{ConstructGraph}_\theta(D_p, \mathcal{O}). \tag{2}$$

Each node is assigned a unique label (e.g., "Postdoctoral Researcher"), an IAD node type, and a set of type-specific attribute values. Each edge is similarly assigned a label (e.g., "funds"), an IG 2.0 edge type, and ABDICO-based attributes. Every

node, edge, and attribute is linked back to its source span in D_p as evidence.

After extraction, TIGRIS applies a normalization and verification step to ensure that each graph element remains within the predefined ontology vocabulary, follows the required schema format, and is supported by evidence spans in D_p. Semantically similar labels are merged using embedding-based similarity to reduce duplicate nodes and rule edges that refer to the same institutional role, action, or policy relation.

3.2 ToC-Guided Policy Impact Simulation

TIGRIS uses G_p to construct policy-grounded agents and organize quantitative estimation along a ToC result chain.

3.2.1 Persona-Grounded Agent Construction

A stakeholder definer first selects actor nodes from G_p for agent construction. For each actor node v, TIGRIS constructs an institutional context K_v by collecting the rule-grounded edges incident to v together with their source evidence, capturing the role-specific obligations, permissions, prohibitions, eligibility conditions, resource access, and execution constraints associated with v. This context serves both as the retrieval scope for persona sampling and the persistent rule memory referenced by the agent during simulation.

To inject role-compatible heterogeneity, TIGRIS translates K_v into a role-conditioned query Q_v over persona attributes (e.g., occupation, education level, age, professional background) and retrieves a candidate pool from the country-specific NVIDIA Nemotron-Personas dataset $\mathcal{N}_{c(p)}$ (Kim et al., 2026; Fujita et al., 2025; Thongpramoon et al., 2026):

$$P_v = \mathrm{ExtractPersona}(\mathcal{N}_{c(p)}, Q_v). \quad (3)$$

Each sampled persona from P_v is bound to v by embedding K_v as structured, persistent fields, providing within-role variation while preserving the agent's institutional reference. Detailed persona fields and filtering criteria are provided in the Appendix C.

3.2.2 ToC Result Chain Construction

Let T_p denote the set of target indicators to be predicted for policy p. TIGRIS operationalizes the ToC using five result-chain phases, $\mathcal{Z} = (\mathrm{In}, \mathrm{Act}, \mathrm{Out}, \mathrm{Ocm}, \mathrm{Imp})$, corresponding to Inputs, Activities, Outputs, Outcomes, and Impacts.

TIGRIS distills two complementary contexts from G_p: a policy context summarizing the policy rules and an organizational context summarizing the participating organizations, jointly denoted as C_p. For each policy p, TIGRIS then constructs a ToC result chain Φ_p that assigns each target indicator $t \in T_p$ to a phase $z \in \mathcal{Z}$ and specifies the discussion agenda and the prediction fields used at that phase:

$$\Phi_p = \mathrm{BuildChain}_\theta(C_p, T_p). \quad (4)$$

The complete phase-by-phase indicator is provided in the Appendix D.

3.2.3 Scenario-Level Simulation

TIGRIS performs scenario-level simulations along the ToC result chain. The *forward direction (Fwd)* follows $\mathcal{Z}$, and the *backward direction (Bwd)* follows the reverse order. A single simulation pass in direction $\delta \in \{\mathrm{Fwd}, \mathrm{Bwd}\}$ denotes one complete traversal over the five phases. Passes can be chained by using the final-phase quantitative estimates and narrative summaries of a preceding pass as the initial context for the next pass. While the forward pass is predictive, the backward pass is abductive, reasoning from the preceding pass's downstream estimates back to upstream conditions consistent with the institutional context. We empirically compare these chain configurations in Sec. 4. Throughout, we use * to denote outputs from the final pass of the chain.

Scenario Setup. For each policy p, a scenario allocator constructs N scenarios by sampling M agents per scenario from the actor-specific pools $\{P_v\}$. We denote the policy-grounded profile of the i-th agent in scenario r as $\rho^i_r \in P_v$.

Per-Phase Procedure. A pass in direction δ orders the five phases as $(z_1, z_2, \ldots, z_5)$. The following procedure is repeated for each z_k, where z_{k-1} denotes the immediately preceding phase (empty when $k = 1$). Each phase consists of two steps:

(i) Agent reasoning. The M agents in scenario r are activated in a randomly sampled order. Given the distilled contexts C_p, the ToC result chain Φ_p, the previous-phase summary $\hat{S}_{r,z_{k-1}}$, and the within-phase discussion history H^i_{r,z_k}, each profile-conditioned agent produces a quantitative target estimate $\hat{\mathbf{e}}^i_{r,z_k}$, a rationale, and evidence references grounded in the institutional context:

$$\hat{\mathbf{e}}^i_{r,z_k} = \mathrm{Agent}_{\theta,\rho^i_r}(C_p, \Phi_p, \hat{S}_{r,z_{k-1}}, H^i_{r,z_k}). \quad (5)$$

(ii) Phase-level aggregation. The agent outputs within the phase are jointly aggregated into a phase-level quantitative prediction $\hat{\mathbf{x}}_{r,z_k}$, computed as the mean of the M agent estimates, and a narrative summary $\hat{S}_{r,z_k}$; together they propagate to the next phase as context:

$$(\hat{\mathbf{x}}_{r,z_k}, \hat{S}_{r,z_k}) = \text{Aggregate}_\theta(\{\hat{\mathbf{e}}^i_{r,z_k}\}_{i=1}^M). \quad (6)$$

Final Prediction. After the chain completes, the per-scenario phase-level outputs are summarized across the N scenarios into a final policy impact report:

$$(\bar{\mathbf{y}}_p, \bar{S}_p) = \text{Summarize}_\theta(\{(\hat{\mathbf{x}}^*_{r,z}, \hat{S}^*_{r,z})\}_{r=1}^N), \quad (7)$$

where $\bar{\mathbf{y}}_p$ contains the per-indicator predictions, computed as the mean of the final-pass estimates at the phase to which each $t \in T_p$ is assigned in Φ_p, and $\bar{S}_p$ is the phase-by-phase impact pathway across all five ToC phases.

4 Experiments

To comprehensively evaluate TIGRIS, we address the following experimental questions:

- **EQ1**: How accurately does TIGRIS predict quantitative policy impact indicators?
- **EQ2**: How do the Policy KG and ToC-guided reasoning contribute to prediction accuracy?
- **EQ3**: Does TIGRIS produce rule-grounded and traceable reasoning that remains institutionally aligned with the policy document?

4.1 Datasets

We evaluate TIGRIS on six policy cases across four countries. The primary evaluation covers four AI talent development policies, and the cross-domain evaluation adds one public-health and one welfare policy.

- AI Singapore 100 Experiments (100E): A project-based AI translation program of Singapore where university researchers and AI Singapore engineers work with industry partners on real-world deployment.
- Brain Korea 21 (BK21): A graduate education and university research-capacity program of Korea that supports advanced research teams in strategic fields, including AI and software.
- Sejong Science Fellowship (SSF): A national fellowship program of Korea that supports early-career STEM researchers, including AI-related postdoctoral researchers, through independent research funding.
- MEXT AIP Project (AIP): A national AI R&D program of Japan that develops AI and big-data research capacity through RIKEN and JST, supporting frontier research and high-level talent development.
- U.S. No Surprises Act: A public-health policy that protects patients from unexpected out-of-network medical bills (Centers for Medicare & Medicaid Services, 2021).
- U.S. Child Care and Development Fund (CCDF) Stabilization Grants: A welfare policy under the American Rescue Plan Act that supported child-care provider operations and workforce retention (Office of Child Care, Administration for Children and Families, 2021).

We refer to these cases as Policies A (100E), B (BK21), C (SSF), D (AIP), E (No Surprises Act), and F (CCDF). We apply the same pipeline to Policies E and F without domain-specific changes. Policies A, B, and E have two quantitative outcome indicators, represented by $\bar{y}_p^{(1)}$ and $\bar{y}_p^{(2)}$, whereas Policies C, D, and F have one. All experiments use only pre-implementation policy documents as simulation input (see Appendix E for details).

4.2 Implementation Details

Baselines. We compare TIGRIS with direct zero-shot prediction, RAG prompting, ToC-only prompting, OASIS (Yang et al., 2024), and SOTOPIA-S4 (Zhou et al., 2025). RAG supplies relevant spans from the same parsed policy document, whereas ToC-only prompting uses the phased result chain without the institutional KG. All methods receive the same policy inputs, target indicators, persona pool, and scenario configuration. To test sensitivity to potentially memorized policy outcomes, we apply identifier masking to every method on Policies B, E, and F by removing policy/program, organization, and geographic names.

LLM Backbones. We implement TIGRIS using three LLM backbones: DeepSeek-V4-Flash (DeepSeek-AI, 2026), GPT-4o-mini (OpenAI, 2024), and Gemini-2.0-Flash (Google DeepMind, 2025). We select these recent instruction-

Table 1: Quantitative prediction errors across the AI Talent policy domain, reported as MAPE. For policies with multiple ground-truth outcome variables, subcolumns denote the components $\bar{y}^{(1)}$ and $\bar{y}^{(2)}$ of $\bar{\mathbf{y}}_p$. The Macro Avg. column reports the mean MAPE across all reported columns. Across all model-method configurations, the lowest and second-lowest reported values in each column are highlighted in **red** and **orange**, respectively.

Dataset (→)		**AI Talent**						**Macro Avg.**
Model (↓)	**Method (↓)**	$\bar{y}_A^{(1)}$	$\bar{y}_A^{(2)}$	$\bar{y}_B^{(1)}$	$\bar{y}_B^{(2)}$	$\bar{y}_C$	$\bar{y}_D$	
GPT-4o-mini	Zero-shot	74.78	79.45	8.27	44.44	39.67	67.29	52.32
	RAG Prompting	74.99	12.42	8.96	18.58	85.73	28.70	38.23
	ToC Prompting	17.04	7.19	14.60	47.52	75.73	74.72	39.47
	SOTOPIA-S4 (Zhou et al., 2025)	37.77	68.88	8.15	21.63	49.67	30.04	36.02
	OASIS (Yang et al., 2024)	47.33	24.72	8.27	31.84	65.62	40.11	36.31
	Ours (*Fwd*)	**7.97**	**1.22**	6.74	**3.47**	27.55	64.87	**18.64**
	Ours (*Fwd*→*Bwd*)	38.15	34.88	7.99	14.19	76.95	59.78	38.66
Gemini-2.0-Flash	Zero-shot	87.83	94.38	8.90	80.53	38.39	70.09	63.35
	RAG Prompting	17.65	22.79	8.99	27.91	68.38	28.70	29.07
	ToC Prompting	25.65	46.23	12.22	40.21	25.58	28.94	29.80
	SOTOPIA-S4 (Zhou et al., 2025)	44.81	54.28	10.41	84.83	68.37	54.28	52.83
	OASIS (Yang et al., 2024)	72.95	83.18	9.51	57.39	78.13	28.92	55.01
	Ours (*Fwd*)	**14.99**	37.83	5.76	43.68	**17.17**	52.60	28.67
	Ours (*Fwd*→*Bwd*)	42.01	37.87	98.91	98.85	52.57	58.10	64.72
DeepSeek-V4-Flash	Zero-shot	74.35	85.21	9.46	35.00	29.92	31.40	44.22
	RAG Prompting	17.65	22.79	8.27	93.06	81.35	**28.68**	41.97
	ToC Prompting	61.96	62.53	8.30	77.70	58.39	29.77	49.77
	SOTOPIA-S4 (Zhou et al., 2025)	44.82	78.21	12.57	52.02	57.03	29.09	45.62
	OASIS (Yang et al., 2024)	46.17	23.68	8.92	43.30	**16.80**	28.70	27.93
	Ours (*Fwd*)	31.71	36.84	**1.77**	**3.97**	35.96	33.31	23.93
	Ours (*Fwd*→*Bwd*)	22.61	24.51	**3.69**	4.47	23.96	**17.67**	**16.15**

Table 2: Cross-domain MAPE under DeepSeek-V4-Flash. Policy E is public health and Policy F is welfare; Macro Avg. is the mean over the three indicators.

Dataset (→)	**Public Health**		**Welfare**	**Macro Avg.**
Method (↓)	$\bar{y}_E^{(1)}$	$\bar{y}_E^{(2)}$	$\bar{y}_F$	
Zero-shot	473.30	647.70	29.20	383.40
RAG Prompting	485.00	650.00	306.67	480.56
ToC Prompting	465.00	577.30	68.30	370.20
SOTOPIA-S4	377.41	511.33	**8.76**	299.17
OASIS	359.36	531.53	140.97	343.95
Ours (*Fwd*)	**1.93**	**6.46**	34.50	**14.30**
Ours (*Fwd*→*Bwd*)	**3.84**	**22.21**	**18.10**	**14.72**

Table 3: Identifier-masking sensitivity under DeepSeek-V4-Flash. Cells report $\Delta\text{MAPE} = \text{MAPE}_{\text{unmasked}} - \text{MAPE}_{\text{masked}}$; negative values indicate higher error after masking. Macro Avg. is the signed mean over five indicators.

Dataset (→)	**AI Talent**		**Public Health**		**Welfare**	**Macro Avg.**
Method (↓)	$\bar{y}_B^{(1)}$	$\bar{y}_B^{(2)}$	$\bar{y}_E^{(1)}$	$\bar{y}_E^{(2)}$	$\bar{y}_F$	
Zero-shot	-0.19	14.17	-56.70	-104.60	-8.30	-31.12
RAG Prompting	0.00	12.35	-3.33	6.82	166.67	36.50
ToC Prompting	0.00	15.50	-1.70	-11.30	10.00	2.50
SOTOPIA-S4	0.11	10.37	15.34	-48.32	-2.88	-5.08
OASIS	-0.54	1.09	-22.24	-19.96	6.67	-7.00
Ours (*Fwd*)	-7.98	0.35	-4.53	3.18	0.05	-1.79
Ours (*Fwd*→*Bwd*)	-4.47	1.40	-6.38	2.70	-1.80	-1.71

tuned models to cover diverse model families and cost–performance profiles for repeated scenario-level simulations. The cross-domain and identifier-masking experiments use DeepSeek-V4-Flash.

Chain Configurations. Building on the *forward (Fwd)* and *backward (Bwd)* reasoning passes defined in Sec. 3.2.3, we empirically compare two chain configurations: *Fwd* and *Fwd*→*Bwd*. These configurations operationalize bidirectional refinement, where the forward-only pass serves as the predictive baseline and the added backward pass tests whether downstream target estimates can re-

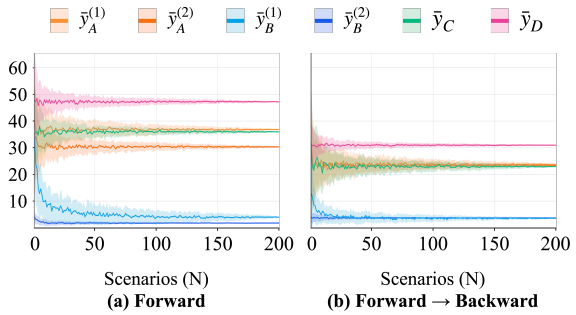


Figure 3: Prediction convergence as scenario count N increases. Lines show mean indicator MAPE; bands are $\pm 1\sigma$ over 500 subsamples drawn without replacement.

vise upstream assumptions and stabilize the resulting impact pathway.

Scenario Setup. For each policy, we generate N = 200 scenarios, each with $M = 5$ agents randomly sampled from the candidate persona set P_v.

5 Results

5.1 Performance of TIGRIS (EQ1)

To answer EQ1, Table 1 compares MAPE on the four primary policies. TIGRIS obtains the lowest macro-average MAPE for each backbone, using *Fwd* for GPT-4o-mini and Gemini-2.0-Flash and *Fwd→Bwd* for DeepSeek-V4-Flash. Its strongest gains occur on Policies A and B; performance is less consistent on Policies C and D, whose indicators reflect broader research productivity. Backward refinement helps DeepSeek-V4-Flash but degrades the other two backbones. Across the representative cases B, E, and F, *Fwd→Bwd* is the only evaluated configuration below 20% policy-level MAPE in all three (4.08, 13.03, and 18.10, respectively), although SOTOPIA-S4 is best on Policy F (Table 2). In the identifier-masking comparison, its policy-averaged MAPE is higher in the masked condition by 1.54, 1.84, and 1.80 points on B, E, and F, respectively, while indicator-level changes vary in direction (Table 3). Figure 3 shows that the mean indicator MAPE becomes broadly stable and the subsampling variance narrows as N increases for the primary Policies A–D.

5.2 Ablation Study (EQ2)

To answer EQ2, we conduct an ablation study on the two core components of TIGRIS, the policy knowledge graph and the ToC-guided simulation. We report the component ablation under the Fwd→Bwd chain as a representative bidirectional setting (Table 4). Relative to the baseline, the Policy KG alone reduces error across most indicators, while ToC alone produces uneven effects—improving some indicators but degrading others. Their combination consistently yields the lowest error across all reported variables, suggesting that institutional grounding from Policy KG and staged reasoning from ToC provide complementary benefits (Table 4).

Policy Document (Policy C)	*This program aims to nurture young scientists in conducting creative, independent research. Approximately 200 fellows are selected annually for a 5-year period (3+2). Eligible applicants are STEM researchers within 7 years of receiving their PhD or under age 39, who are not tenure-track university faculty; only Korean nationals may apply during the first 3 years of the program. The fellowship is designed for solo research only; [...]*
TIGRIS (Ours)	*"As a member of the Basic Research Project Promotion Committee, I see the fellowship as a highly attractive opportunity for early-career researchers, given the stable salary and research funding. I estimate about 1,500 applications, reflecting strong interest but also the restrictive eligibility (domestic, non-tenure-track, within 7 years of PhD or under 39). The policy explicitly states 200 fellows selected annually [...] as the policy allows one change per year, but most will stay due to stability concerns."*
SOTOPIA S4	*"Based on my survey of 27 K-12 schools in Seodaemun-gu, Seoul, an entirely different pattern emerges: the average SCI(E) papers per research was 0.3 in 2021, dropping 40% to 0.18 in 2022, even though research-data-management-system adoption in schools rose from 67% to 89% over the same period. [...] exactly matching the drop in paper productivity."*
OASIS	*"Given that each fellow is a highly productive early-career researcher, a conservative estimate of 3-4 SCI(E) papers per fellow per year is reasonable. Thus, 200 fellows × 3 papers/year yields about 600 papers per year, but considering the [...] as well as the inclusion of 50 additional linked fellows, I estimate an annual output of approximately 1,200 SCI(E) papers} for the [...]"*

Figure 4: Examples of reasoning generated by TIGRIS and the baselines on Policy C. indicates policy-grounded statements, and indicates policy deviations.

Because the ablation results suggest that the Policy KG contributes to TIGRIS's gains, we further validate whether the extracted graphs faithfully represent document-specified institutional structure. The validation shows that explicit institutional entities and evidence links are comparatively reliable, whereas implicit conditions, execution constraints, and exception clauses are harder to recover. These findings support the Policy KG as an auditable grounding layer for simulation, while clarifying that it is not a fully verified formal constraint system; full details are provided in Appendix J.

5.3 Qualitative Analysis (EQ3)

To answer EQ3, we evaluate reasoning faithfulness because policy impact simulation has no single ground-truth reasoning path. On the same randomly sampled 10% subset of Policy C used for qualitative analysis, five annotators indepen-

Table 4: Component ablation under DeepSeek-V4-Flash with the *Fwd→Bwd* chain. The Baseline row denotes the configuration without either the Policy KG or ToC. Cells report MAPE, and the Macro Avg. column reports the mean MAPE across all reported indicators. The lowest reported values are highlighted in **red**.

Method	Components		AI Talent						Macro Avg.
	Policy KG	ToC	$\bar{y}_A^{(1)}$	$\bar{y}_A^{(2)}$	$\bar{y}_B^{(1)}$	$\bar{y}_B^{(2)}$	$\bar{y}_C$	$\bar{y}_D$	
Baseline	✗	✗	92.32	96.25	11.00	43.83	95.09	526.50	144.17
Ours	✓	✗	72.44	85.38	5.49	25.24	26.30	65.20	46.68
	✗	✓	69.24	81.63	8.58	90.54	25.30	76.58	58.65
	✓	✓	**22.61**	**24.51**	**3.69**	**4.47**	**23.96**	**17.67**	**16.15**

dently assessed 60 utterances per method for evidence support, hallucination, rule violation, and coherence. TIGRIS achieved the highest evidence-support score (1.3/2) and coherence score (4.3/5), together with the lowest hallucination (3.3%) and rule-violation (0.0%) rates. Appendix I reports the full protocol, method-level results, and inter-annotator agreement, while Figure 4 presents a representative example from Policy C.

In contrast, the baselines exhibit two distinct patterns of institutional inconsistency. SOTOPIA-S4 (Zhou et al., 2025) shows statistical hallucination, fabricating unsupported statistics and arbitrary variables that lack any basis in the policy document or verifiable sources. OASIS (Yang et al., 2024) exhibits both rule violation and population conflation, incorrectly merging researchers from a separately administered training program into the Policy C beneficiary pool. A complementary case study tracing this institutional grounding across the full ToC chain is provided in Appendix G.

6 Conclusion

TIGRIS is an end-to-end framework that converts policy documents into an IAD/IG 2.0-based knowledge graph and anchors policy-grounded persona agents to a Theory of Change for causal reasoning. Across four primary AI talent development policies and three LLMs, TIGRIS achieves the lowest macro-average MAPE while producing rationales traceable to document-derived rules and explicit causal assumptions. Additional evaluations on U.S. public-health and welfare policies provide initial evidence of cross-domain applicability without domain-specific pipeline changes. Beyond accuracy, these results suggest a path for NLP to evolve from open-ended generation toward theory-grounded, auditable reasoning in high-stakes institutional domains—including AI talent development, public health, and welfare—supporting transparent, stakeholder-informed pre-implementation policy decisions.

Limitations

Our evaluation covers four AI talent development policies from three countries and two U.S. cross-domain cases, one in public health and one in welfare, evaluated with DeepSeek-V4-Flash. These cases provide initial evidence beyond AI talent policy but do not establish generalizability across domains, countries, or backbones. Backward refinement reduces error with DeepSeek-V4-Flash but increases it with GPT-4o-mini and Gemini-2.0-Flash, making it an optional, backbone-dependent component. Future work should evaluate more policies and backbones across climate, industrial, and education policy and compare TIGRIS with agent-based, forecasting, and causal methods where data permit.

Because TIGRIS relies on automatically extracted policy KGs and LLM-based reasoning, failures to recover implicit conditions, execution constraints, or exception clauses may propagate through the ToC pathway, as discussed in Appendix J. Future work should assess sensitivity to KG construction, ToC design, and simulation assumptions. TIGRIS should therefore support, rather than replace, policy expertise through expert review, sensitivity analysis, and transparent uncertainty reporting.

Acknowledgments

This work was supported by grants from the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT) (No. RS-2020-II201373, Artificial Intelligence Graduate School Program, Hanyang University; No. RS-2025-25464239, Research on Securing National AI Core and High-Level Human Resources; and No. RS-2026-25548925).

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Centers for Medicare & Medicaid Services. 2021. Requirements related to surprise billing; part i interim final rule with comment period. Published July 1, 2021.

Qi Chai, Zhang Zheng, Junlong Ren, Deheng Ye, Zichuan Lin, and Hao Wang. 2025. CausalMACE: Causality empowered multi-agents in Minecraft cooperative tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14410–14426, Suzhou, China. Association for Computational Linguistics.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Sue E. S. Crawford and Elinor Ostrom. 1995. A grammar of institutions. *American Political Science Review*, 89(3):582–600.

Sharon S Dawes and Marijn Janssen. 2013. Policy informatics: addressing complex problems with rich data, computational tools, and stakeholder engagement. In *Proceedings of the 14th annual international conference on digital government research*, pages 251–253.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report and model card for DeepSeek-V4-Pro and DeepSeek-V4-Flash.

Christopher K. Frantz and Saba Siddiki. 2021. Institutional grammar 2.0: A specification for encoding and analyzing institutional design. *Public Administration*, 99(2):222–247.

Atsunori Fujita, Vincent Gong, Masaya Ogushi, Kotaro Yamamoto, Yoshi Suhara, Dane Corneil, and Yev Meyer. 2025. Nemotron-Personas-Japan: Synthetic personas aligned to real-world distributions.

Sue C. Funnell and Patricia J. Rogers. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. John Wiley & Sons.

Google DeepMind. 2025. Gemini 2.0 flash model card. Model card published April 15, 2025.

Dongyoon Hahm, Woogyeol Jin, June Suk Choi, Sungsoo Ahn, and Kimin Lee. 2025. Enhancing llm agent safety via causal influence prompting. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15143–15168.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling (COLM)*.

Jiarui Ji, Yang Li, Hongtao Liu, Zhicheng Du, Zhewei Wei, Qi Qi, Weiran Shen, and Yankai Lin. 2024. SRAP-agent: Simulating and optimizing scarce resource allocation policy with LLM-based agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 267–293, Miami, Florida, USA. Association for Computational Linguistics.

Erik W. Johnston, editor. 2015. *Governance in the Information Era: Theory and Practice of Policy Informatics*. Routledge, New York.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. Llms are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569.

Hyunwoo Kim, Jihyeon Ryu, Jinho Lee, Hyungon Ryu, Kiran Praveen, Shyamala Prayaga, Kirit Thadaka, Will Jennings, Bardiya Sadeghi, Ashton Sharabiani, Yejin Choi, and Yev Meyer. 2026. Nemotron-personas-korea: Synthetic personas aligned to real-world distributions for korea.

Hao Li, Ruoyuan Gong, and Hao Jiang. 2025. Political actor agent: Simulating legislative system for roll call votes prediction with large language models. 39(1):388–396.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Michael D. McGinnis. 2011. An introduction to iad and the language of the ostrom workshop: A simple guide to a complex framework. *Policy Studies Journal*, 39(1):169–183.

Xinyi Mou, Zejun Li, Hanjia Lyu, Jiebo Luo, and Zhongyu Wei. 2024a. Unifying local and global knowledge: Empowering large language models as political experts with knowledge graphs. In *Proceedings of the ACM Web Conference 2024*, pages 2603–2614.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024b. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4789–4809, Bangkok, Thailand. Association for Computational Linguistics.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. 2024. Can LLMs follow simple rules? *arXiv preprint arXiv:2311.04235*.

Taehyung Noh, Seungwan Jin, Haein Yeo, and Kyungsik Han. 2025. Externalizing social-cognitive structures for user modeling: Toward theory-driven profiling with llms. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5063–5067.

Taehyung Noh, Seungwan Jin, Haein Yeo, and Kyungsik Han. 2026. Triple: Theory-driven integration of planned and habitual behaviors for llm-based personalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17626–17634.

OECD. 2022. Glossary of key terms in evaluation and results-based management for sustainable development. Technical report, OECD Publishing, Paris.

Office of Child Care, Administration for Children and Families. 2021. ARP act child care stabilization grants. Information Memorandum CCDF-ACF-IM-2021-02, published May 10, 2021.

OpenAI. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Elinor Ostrom. 2005. *Understanding Institutional Diversity*. Princeton University Press, Princeton, NJ.

Elinor Ostrom. 2011. Background on the institutional analysis and development framework. *Policy Studies Journal*, 39(1):7–27.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36:3580–3599.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Pongsasit Thongpramoon, Verdi March, Christopher Low, Shyamala Prayaga, Dane Corneil, and Yev Meyer. 2026. Nemotron-Personas-Singapore: Synthetic Personas Aligned to Real-World Distributions for Singapore.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. *arXiv preprint arXiv:2312.03664*.

Carol H Weiss and 1 others. 1995. Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. *New approaches to evaluating community initiatives: Concepts, methods, and contexts*, 1:65–92.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, and 1 others. 2024. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jentse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, Hao Zhu, and Maarten Sap. 2025. SOTOPIA-s4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 350–360, Albuquerque, New Mexico. Association for Computational Linguistics.

A Algorithmic Details

Algorithm 1 summarizes the end-to-end execution of TIGRIS.

Algorithm 1 TIGRIS

Require: Policy PDF P_p, ontology schema $\mathcal{O}$, target indicators T_p, country-specific persona pool $\mathcal{N}_{c(p)}$, ToC phase set Z, number of scenarios N, agents per scenario M, chain configuration Δ
Ensure: Target prediction $\bar{\mathbf{y}}_p$ and final impact report $\bar{\mathcal{S}}_p$

procedure TIGRIS($P_p, \mathcal{O}, T_p, \mathcal{N}_{c(p)}, Z, N, M, \Delta$)

Stage 1. Institutional Policy Knowledge Graph Construction
$D_p \leftarrow$ PARSEPOLICYPDF(P_p)
$\mathcal{U}_p \leftarrow$ SEGMENTDOCUMENT(D_p)
$V_p \leftarrow \emptyset;\ E_p \leftarrow \emptyset;\ \eta_p \leftarrow \emptyset$
for all $u_j \in \mathcal{U}_p$ **do**
  $N_j \leftarrow$ LLMNODEEXTRACT($u_j, \mathcal{O}$) ▷ # IAD-based nodes
  $R_j \leftarrow$ LLMRULEEXTRACT($u_j, V_p \cup N_j, \mathcal{O}$) ▷ # IG 2.0 rule edges
  $V_p \leftarrow V_p \cup N_j;\ E_p \leftarrow E_p \cup R_j$
  $\eta_p \leftarrow$ UPDATEEVIDENCE($\eta_p, N_j \cup R_j, u_j$)
end for
$(V_p, E_p) \leftarrow$ NORMALIZEANDVERIFY($V_p, E_p, \eta_p, \mathcal{O}$) ▷ # Embedding-based merge
$G_p \leftarrow$ BUILDGRAPH(V_p, E_p, η_p)

Stage 2. ToC-Guided Policy Impact Simulation
$\mathcal{V}_p^A \leftarrow$ SELECTACTORNODES(G_p)
for all $v \in \mathcal{V}_p^A$ **do**
  $K_v \leftarrow$ COLLECTRULECONTEXT(v, G_p, η_p)
  $Q_v \leftarrow$ BUILDQUERY(K_v)
  $P_v \leftarrow$ EXTRACTPERSONA($\mathcal{N}_{c(p)}, Q_v$) ▷ # Eq. (3)
end for
$C_p \leftarrow$ DISTILLCONTEXT(G_p) ▷ # Policy + organizational contexts
$\Phi_p \leftarrow$ BUILDCHAIN(C_p, T_p) ▷ # Eq. (4)
for $r = 1$ **to** N **do**
  $\mathcal{A}_r \leftarrow$ SAMPLEAGENTS($\{P_v\}_{v \in \mathcal{V}_p^A}, M$)
  $\mathcal{X}_r^{\text{prev}} \leftarrow \varnothing$
  for all ordered $\delta \in \Delta$ **do**
    $\mathcal{X}_r^{\delta} \leftarrow$ RUNTOCPASS($\mathcal{A}_r, C_p, \Phi_p, Z, \delta, \mathcal{X}_r^{\text{prev}}$)
    $\mathcal{X}_r^{\text{prev}} \leftarrow \mathcal{X}_r^{\delta}$
  end for
  $\mathcal{X}_r^{*} \leftarrow \mathcal{X}_r^{\text{prev}}$ ▷ # $*$ = final pass
end for
$(\bar{\mathbf{y}}_p, \bar{\mathcal{S}}_p) \leftarrow$ SUMMARIZE($\{\mathcal{X}_r^{*}\}_{r=1}^{N}$) ▷ # Eq. (7)
return $\bar{\mathbf{y}}_p, \bar{\mathcal{S}}_p$
end procedure

procedure RUNTOCPASS($\mathcal{A}, C_p, \Phi_p, Z, \delta, \mathcal{X}^{\text{prev}}$)
$(z_1, \ldots, z_5) \leftarrow$ ORDERPHASES(Z, δ)
$\hat{S}_{z_0} \leftarrow$ TAILSUMMARY($\mathcal{X}^{\text{prev}}$) ▷ # Carry-over context
for $k = 1$ **to** 5 **do**
  for all $\rho^i \in \mathcal{A}$ in random order **do**
    $\hat{e}_{z_k}^i \leftarrow$ AGENT($C_p, \Phi_p, \hat{S}_{z_{k-1}}, H_{z_k}^i; \rho^i$) ▷ # Eq. (5)
  end for
  $(\hat{x}_{z_k}, \hat{S}_{z_k}) \leftarrow$ AGGREGATE($\{\hat{e}_{z_k}^i\}_{i=1}^{M}$) ▷ # Eq. (6)
end for
return $\{(\hat{x}_{z_k}, \hat{S}_{z_k})\}_{k=1}^{5}$
end procedure

B Prompt Templates

B.1 Node Prompt

Figure 5 and Figure 6 illustrate how the IAD-based node schema are operationalized in the prompt templates for policy knowledge graph construction.

KG Node Extractor System

You are an expert policy knowledge graph extractor for IAD/IG-based policy impact simulation. Respond only in valid JSON with no markdown formatting or code blocks. All generated node labels, summaries, attributes, warnings, and source_text evidence must be in English.

Figure 5

KG Node Extractor User

Extract policy KG node instances from the document chunk using the provided IAD/IG ontology.

Rules:
- Use only ontology-defined entity types and keep node names, summaries, attributes, and source evidence grounded in the document.
- Extract both stakeholder actors and non-actor policy objects when they explain actor constraints, information, actions, costs, benefits, authorities, or outcomes.
- Treat persona-eligible actors, roles, organizations, and positions as simulation actor anchors rather than individual biographies.
- For each persona-eligible node, specify agent_modality and role_family to guide later persona generation and policy simulation.
- Use node_family to ensure coverage of actors, roles, eligibility conditions, restrictions, processes, information evidence, resources, authorities, milestones, and outcomes.
- Assign each extracted node an IAD rule type from the ontology: position, boundary, choice, aggregation, information, payoff, or scope.
- Extract conditions, restrictions, process steps, dates, decision authorities, evaluation criteria, resource components, and outcome metrics as separate node instances when they can affect simulation behavior.
- Fill attributes only with grounded information useful for policy simulation or persona synthesis, such as eligibility criteria, role scope, decision authority, support period, funding amount, or evaluation criteria.
- Do not invent missing eligibility criteria, actor motives, demographic details, or unsupported policy relationships.
- Target a compact, edge-ready node set by keeping nodes that can support downstream relations and simulation behavior.

Figure 6

B.2 Edge Prompt

Figure 7 and Figure 8 illustrate how the IG 2.0-based rule schema are operationalized in the prompt templates for policy knowledge graph construction.

KG Edge Extractor System

You are an expert policy knowledge graph edge extractor using IAD (Ostrom) and IG 2.0 (Institutional Grammar 2.0) for policy impact simulation.

First classify each candidate institutional statement:
regulative = an actor must/can/cannot do an action to an object.
constitutive = a term, condition, action, object, or status is defined as, means, counts as, constitutes, or qualifies as another institutional category.
structural/strategy = a process, resource flow, evaluation relation, information relation, or causal relation without deontic or definition function.

For regulative statements, decompose using IG 2.0 ABDICO syntax:
A = Attribute (the actor who holds the rule — maps to edge.source)
D = Deontic (must / can / cannot — maps to edge.deontic)
I = aIm (the action verb — maps to edge.type)
B = oBject (the entity acted upon — maps to edge.target)
C = Conditions, split into TWO IG 2.0 sub-fields:
Cac (Activation Condition) — when does the rule apply / who is in scope (eligibility, target group, triggering event, time window). Maps to edge.activation_condition.
Cex (Execution Constraint) — how must the action be performed once the rule fires (manner, deadline within the rule, format, prerequisite approval, frequency limit). Maps to edge.execution_constraint.
O = Or-else (sanction or consequence on non-compliance — maps to edge.or_else_text)

For constitutive statements, do not force ABDICO. Extract the defined term or qualifying condition as source, the institutional category/status as target, set edge_category="constitutive", deontic=null, and use a DefinitionQualification edge type when the ontology provides one.

Respond only in valid JSON with no markdown formatting or code blocks. All generated facts, conditions, attributes, warnings, and source_text evidence must be in English.

Figure 7

KG Edge Extractor User

Extract policy KG edge instances from the document chunk using the provided ontology and global node pool.

- Use only ontology-defined edge types, and ensure that every source and target is an exact node ID from the Global Node Pool.
- Extract edges only when the source text directly supports the relation, including the edge type, source, target, deontic relation, and conditions when present.

- Extract structural relations for process order, eligibility, selection, resource flow, implementation, evaluation, submission, reporting, monitoring, and information use.
- Extract constitutive IG relations for explicit definition, meaning, counts-as, constitutes, or qualification clauses.
- Prefer specific, source-supported relations that can change downstream simulation behavior.

- Classify each edge according to IG 2.0 statement types:
  - Regulative: institutional rules about what an actor must, can, or cannot do.
  - Constitutive: definitions of what something means, counts as, constitutes, or qualifies as.
  - Structural: factual process, flow, evaluation, implementation, or causal relations.

- For regulative edges, apply ABDICO decomposition:
  - A: identify the actor, role, organization, or position that holds the rule.
  - D: assign the deontic relation as must, can, or cannot.
  - I: map the institutional action to an ontology-defined edge type.
  - B: identify the policy object, document, resource, role, decision, or process affected by the action.
  - C: split conditions into activation_condition and execution_constraint.
  - O: extract explicit sanctions or consequences as or_else_text.

- Separate activation conditions from execution constraints:
  - activation_condition defines when the rule applies or who is in scope.
  - execution_constraint defines how the action must be performed once the rule applies.

- Regulative rules: institutional statements about what an actor must, can, or cannot do.
- Non-regulative relations: constitutive definitions and structural process, flow, evaluation, implementation, or causal relations.

Figure 8

C Persona Fields and Filtering Criteria

Table 5 lists the persona attribute fields used during retrieval and agent construction. We use a compact subset of the full Nemotron-Personas schema, restricted to fields that affect institutional role.

Field	Use in TIGRIS
`occupation`	Role match against IAD position (e.g., researcher)
`education_level`	Eligibility filter (e.g., Doctorate, Undergraduate)
`age`	Age-based eligibility filter
`country`, `city`	Match against $c(p)$
`professional_persona`	Narrative background for prompt context
`skills_and_expertise`	Domain match (e.g., AI/ML/policy)
`career_goals`	Within-role heterogeneity
`sex`	Reported only; not used for filtering

Table 5: Persona fields used by TIGRIS.

C.1 Scope Query Generation

For each actor node v, the stakeholder definer converts the institutional context K_v into a tiered scope query

$$Q_v = (Q_v^{\text{dir}}, Q_v^{\text{pla}}, Q_v^{\text{wk}}), \tag{8}$$

where Q_v^{dir}, Q_v^{pla}, and Q_v^{wk} denote direct, plausible, and weak role-fit filters with decreasing institutional fit. Each filter is defined over Nemotron fields, including age, education_level, occupation, and domain fields when available. The stakeholder definer selects filter values from the available value lists in the country-specific Nemotron pool $\mathcal{N}_{c(p)}$, rather than generating free-form labels.

C.2 Persona Retrieval

The tiered query is executed as SQL filters over $\mathcal{N}_{c(p)}$. For each personal-level actor node, TIGRIS retrieves up to 300 candidate personas. Retrieval first applies the direct tier. If fewer than 300 candidates are returned, the plausible tier is applied and appended after deduplication. If the pool is still smaller than 300 candidates, the weak tier is applied as the final fallback. The ordered retrieval process is written as

$$P_v = \text{Top}_{300}\left(R_v^{\text{dir}} \oplus R_v^{\text{pla}} \oplus R_v^{\text{wk}}\right), \tag{9}$$

where $R_v^t = \text{SQLRetrieve}(\mathcal{N}_{c(p)}, Q_v^t)$ for tier t, and $\oplus$ denotes ordered concatenation after deduplication.

Persona retrieval is applied to actor nodes that are instantiated as personal-level agents. When an organizational actor must participate in the simulation, TIGRIS uses the associated institutional position or representative role as the retrieval anchor.

D ToC Schema

Figure 9 illustrates the ToC schema adopted in the TIGRIS simulation framework, encompassing Inputs, Activities, Outputs, Outcomes, and Impacts. The schema was implemented using the latest experimental model, DeepSeek-V4-Pro, to support structured and consistent reasoning throughout the simulation and evaluation process.



Figure 9: Examples of Theory of Change Schema built for simulation process in Policy A.

E Dataset Details

	Policy A (100E)	**Policy B (BK21)**	**Policy C (SSF)**	**Policy D (AIP)**
Country	Singapore	Korea	Korea	Japan
Operating Body	AI Singapore (AISG), under NRF Singapore.	NRF Korea, Ministry of Education.	National Research Foundation (NRF) of Korea.	RIKEN AIP and Japan Science and Technology Agency (JST), under Japan's Ministry of Education (MEXT).
Purpose & Mechanism	Project-based AI translation program of Singapore. University researchers and AI Singapore engineers work with industry partners on real-world AI deployment, with co-funding up to SGD 150,000 per project over 6–18 months.	Graduate education and university research-capacity program of Korea ("BK21 FOUR") supporting advanced research teams in strategic fields, including AI and software. Reports training-outcome and research-collaboration indicators at the university level.	National fellowship program of Korea supporting early-career STEM researchers, including AI-related postdoctoral researchers, through independent research funding. Selects approximately 200 fellows annually for a 3+2-year period.	National advanced-AI R&D program of Japan that develops AI, big-data, and related research capacity through RIKEN and JST, supporting frontier AI research and high-level talent development.
Target Population	Industry partners jointly with university researchers and AISG engineers.	Graduate students and faculty in BK21-funded Education and Research Groups at Korean universities.	STEM researchers within 7 years of PhD or under age 39; excludes tenure-track faculty; Korean nationals only during the first 3 program years.	AIP-affiliated principal investigators, postdoctoral researchers, doctoral students, and partner institutions under MEXT.
Outcome Indicator(s)	(1) Approved 100E projects (cumulative). (2) Deployed 100E projects (cumulative).	(1) Trainee employment rate (%). (2) International collaboration rate (%).	SCI(E)-indexed publications.	Top-conference papers.
Reporting Period	2023.	2022.	2021–2022.	April 2021 – March 2022.
Ground Truth	(1) 115 (2) 73	(1) 82.2 (2) 36.0	721	108
Source	(1) AISG 5-Year Report; (2) Singapore Digital Economy Report 2024 by the Infocomm Media Development Authority.	BK21 FOUR Annual Report, Hanyang University.	NRF Performance Report (cohort of 648 active fellows).	RIKEN AIP Annual Report.

Table 6: Four AI talent development policies used for evaluation. Policies A and B contribute two outcome indicators each, denoted (1) and (2), which correspond to the components $\hat{y}_p^{(1)}$ and $\hat{y}_p^{(2)}$.

F Ontology Details

This section details the ontology used to construct the policy knowledge graph. It follows the schema in Sec 3.1, where node types are derived from the Institutional Analysis and Development (IAD) framework, edge types are derived from Institutional Grammar 2.0 (IG 2.0), and attributes store policy-specific constraints such as eligibility, support period, activation condition, execution constraint, and source evidence. The goal is not to build a general entity-relation graph, but to build a compact institutional rule memory that can be referenced throughout ToC-guided simulation.

Symbol	Schema part	Definition in TIGRIS	Stored with
T_{IAD}	IAD node vocabulary	Institutional entities used to represent who acts, what role they hold, what action or information is relevant, and which cost or benefit is attached to the policy.	Node type and node attributes.
T_{IG}	IG 2.0 edge vocabulary	Rule-grounded relation labels used to encode policy actions, authority relations, definitions, resources, and ToC pathway links.	Edge type and rule attributes.
A	Attribute schema	Policy-specific values beyond labels and types, including eligibility, amount, time period, condition, sanction, and evidence.	Node and edge attribute maps.
η	Evidence map	Source-span references that justify each extracted node, edge, or attribute.	Extraction record and agent context.

Table 7: Global ontology schema used by TIGRIS. The schema separates vocabulary, attributes, and evidence so that simulation agents can reference policy rules without treating the full document as unconstrained prompt context.

F.1 IAD Node Vocabulary

Node type	Definition	Use in simulation
Participant	Individual or organizational actor appearing in implementation, funding, oversight, participation, or evaluation.	Defines the actor universe. Examples include ministries, agencies, universities, research groups, enterprises, programme operators, and beneficiaries.
Position	Institutional role assigned to a participant. A position defines policy authority or obligation, not demographic identity.	Primary anchor for persona construction. Rules incident to a position become the agent's persistent institutional context.
Action	Policy-defined behavior, administrative step, or implementation operation.	Specifies what can occur in the policy pathway, such as apply, submit, evaluate, allocate, train, deploy, monitor, or report.
Information	Document, criterion, metric, form, guideline, report, or knowledge resource used during implementation.	Grounds reasoning in accessible policy information, including evaluation criteria, application guidelines, reporting templates, and target indicators.
Costs and benefits	Resource, cost, support package, reward, sanction, exclusion, refund, or other consequence.	Represents budgets, grants, co-funding caps, stipends, salary support, audits, refunds, and participation restrictions.

Table 8: IAD-based node vocabulary. Participant and Position nodes form the actor-node subset used for persona construction, while the remaining nodes provide the institutional context referenced across ToC phases.

F.2 IG 2.0 Rule-Edge Vocabulary

TIGRIS separates normalized edge labels from rule attributes. For regulative statements, the edge label records the policy action, while the deontic modality is stored as an ABDICO attribute. This prevents the graph from creating separate labels for the same action under different modalities. For example, an actor may be required, permitted, or prohibited to submit a report; all three cases use the same normalized action edge and differ in the stored deontic value.

Edge class	Normalized labels	Stored attributes and simulation use
Regulative action	Apply to; submit; conduct; receive support; change host; report; evaluate; select; terminate.	Stores ABDICO fields: actor, deontic, aim, object, activation condition, execution constraint, and or-else consequence. These edges define what an agent must, can, or cannot do.
Constitutive definition	Defines as; counts as; qualifies as; has status; has eligibility criterion; has target metric.	Stores term scope, category, qualifying condition, indicator definition, and source evidence. These edges resolve policy-specific categories before simulation.
Resource and authority	Governs; funds; allocates; supports; monitors; approves; audits.	Stores authority scope, amount, currency, frequency, reporting line, and evidence. These edges define budget flow, oversight, accountability, and control.
Implementation pathway	Implements; collaborates with; produces; contributes to; requires; depends on.	Stores ToC phase, pathway role, expected output, target indicator, and condition. These edges link institutional actions to Inputs, Activities, Outputs, Outcomes, and Impacts.

Table 9: IG 2.0-based edge vocabulary. Edge labels stay compact, while rule-specific information is stored as structured attributes.

Field	Name	Meaning in TIGRIS	Extraction constraint
A	Attribute	Actor or institutional position to which the rule applies.	Must map to a Participant or Position node when the source span identifies an actor or role.
D	Deontic	Normative modality, such as must, can, or cannot.	Filled for regulative statements; left empty for purely definitional or structural relations.
I	Aim	Policy action represented by the normalized edge label.	Must be action-centered, such as submit, evaluate, allocate, train, deploy, or report.
B	Object	Entity, document, resource, decision, or process affected by the action.	Mapped to a node when possible; otherwise stored as a normalized object attribute with evidence.
C-act	Activation condition	Conditions under which the rule becomes applicable.	Captures eligibility, target population, geography, time period, track, status, or prerequisite.
C-ex	Execution constraint	Conditions governing how the action must be performed once active.	Captures deadline, frequency, amount, format, procedure, continuation rule, or reporting requirement.
O	Or else	Sanction, exclusion, termination, refund, non-selection, or other consequence.	Stored only when the source span states or directly implies a consequence for non-compliance.

Table 10: ABDICO fields stored on rule-grounded edges. TIGRIS separates activation conditions from execution constraints because the former define when a rule governs an actor, while the latter define how the governed actor must act.

F.3 ABDICO Attribute Schema

F.4 Illustrative Rule Decomposition

Component	Illustrative value for a fellowship eligibility rule
Source rule	Early-career researchers may apply only when they satisfy the programme eligibility scope, including degree-age conditions, appointment-status conditions, nationality or residency rules, and the announced application period.
Edge class and label	Regulative action edge: apply to.
A: Attribute	Early-career researcher, represented as a Position node linked to eligible applicant participants.
D: Deontic	can.
I: Aim	Apply to the fellowship or programme.
B: Object	Fellowship application.
C-act: Activation condition	Applicant falls inside the stated eligibility scope when the call is active.
C-ex: Execution constraint	Application follows the announced proposal, submission, and evaluation procedures.
O: Or else	Applicants outside the eligibility scope are excluded from the eligible applicant pool.
Simulation use	Restricts persona retrieval and prevents agents from treating unrelated student-training, school-level, or organization-only actors as eligible beneficiaries.

Table 11: Illustrative ABDICO decomposition. The example shows the format of the rule memory supplied to agents; it is not used as a substitute for the full policy document.

F.5 Normalization and Evidence Rules

After extraction, TIGRIS applies four normalization rules. First, synonymous role labels are merged only when they refer to the same institutional position within the same policy. Second, organization nodes are not merged with role nodes because one organization can contain several positions with different authority, obligation, and resource-access rules. Third, a rule edge is retained only when its actor, action, object, and condition are supported by source evidence. Fourth, duplicate rule edges are merged by keeping the most specific condition and aggregating evidence spans, so repeated passages do not overweight a rule during simulation.

G Case Study

This section provides a compact qualitative case study for Policy A, AI Singapore 100 Experiments (100E). Rather than presenting a single cherry-picked trajectory, we compare two randomly sampled saved scenario logs, `S10` and `S16`. Each log contains five persona agents sampled from policy-grounded role pools and follows the same Inputs–Activities–Outputs–Outcomes–Impacts ToC structure. The values in Table 12 are scenario-level examples used to inspect traceability; they are not substitutes for the aggregate predictions reported in the main experiment.

Log and agents	Forward feasibility trace	Final-chain target-pressure trace	Diagnostic reading
`S10`: two AI practitioners, two AI researchers, and one Smart Nation policy-role agent.	Inputs are stable: S$500M funding, S$330K co-funding cap, 500K GPU hours, and 50 trainers. The forward path yields 40 outreach sessions, 20 matchmaking events, 20 workshops, 35 onboarding assists, 35 approved projects, 12 deployed projects, and 12% enterprise adoption.	The final F-B-F-B chain raises the required scale to 90 approved projects, 70 deployed projects, and 30% adoption, with impact estimates of rank 8, 3.5% AI GDP share, trust score 65, and 70% retention.	Conservative first-year rollout. The same policy anchors support a moderate forward pipeline, but the backward pass exposes how much more activity and deployment would be required to satisfy stronger outcome targets.
`S16`: one AI practitioner, three AI researchers, and one Smart Nation policy-role agent.	Inputs remain identical: S$500M funding, S$330K co-funding cap, 500K GPU hours, and 50 trainers. The forward path yields 29 outreach sessions, 14.4 matchmaking events, 19.6 workshops, 48 onboarding assists, 80 approved projects, 50 deployed projects, and 22% enterprise adoption.	The final F-B-F-B chain retains a higher-output pathway: 78 approved projects, 58 deployed projects, and 25% adoption, with impact estimates of rank 6, 3.74% AI GDP share, trust score 61.6, and 71.6% retention.	Higher-conversion sampled trajectory. The institutional anchors are unchanged, but the agents assume a more effective translation from enterprise engagement to deployment; the report should still preserve caveats about production integration and enterprise readiness.

Table 12: Compact case-study comparison across two randomly sampled Policy A logs. F-B-F-B abbreviates the Fwd→Bwd→Fwd→Bwd chain. The table integrates setup, ToC trace, and interpretation into one view to avoid unnecessary subsectioning.

The comparison shows that TIGRIS makes two different sources of variation visible. Policy-derived anchors such as funding, co-funding, compute, and trainer capacity remain stable across sampled logs. In contrast, downstream quantities vary when agents infer conversion rates from outreach to approvals, from approved projects to deployed projects, and from deployment to enterprise adoption. The case study therefore illustrates the main purpose of the framework: the policy knowledge graph constrains institutionally specified quantities, while the ToC chain exposes where causal assumptions drive the final estimate.

H Examples of Final Report

TIGRIS generates a final report after a simulation chain is completed. The report is not a transcript of all agent messages; it is an audit-oriented summary containing the final-pass target estimates, the phase pathway that produced them, and the main residual caveat. Table 13 shows compact report-style extracts from three randomly sampled saved logs. These rows illustrate the report format only. In the main experiment, the policy-level prediction is computed by averaging the final-pass estimates across the full scenario pool.

Sampled report	Final-pass target estimate	Pathway evidence retained in the report	Residual caveat
Policy B / `S9`	Graduate employment rate: 72%. International collaboration rate: 35%.	The final backward chain links these outcomes to 12,600 supported graduate students, 6,400 postdoctoral researchers, 12,000 publications, and 4,000 international conference presentations.	Employment and collaboration should not be interpreted as driven by the same mechanism. Employment is tied to training and programme reputation, whereas collaboration depends on stronger co-authorship and network-building assumptions.
Policy A / `S10`	Deployed 100E projects: 70. Enterprise AI adoption rate: 30%.	The report keeps the stable input anchors of S$500M funding, S$330K co-funding cap, 500K GPU hours, and 50 trainers, then records that the final chain requires 90 approved projects to support 70 deployments.	This is a target-pressure report rather than a purely forward-feasible first-year estimate. The large gap between the forward deployment estimate and the final-chain requirement should be reported as deployment risk.
Policy A / `S16`	Deployed 100E projects: 58. Enterprise AI adoption rate: 25%.	The final chain retains the same input anchors but assumes a stronger conversion pathway than `S10`: 78 approved projects, 58 deployed projects, and 25% adoption, with impact estimates of rank 6, 3.74% AI GDP share, trust score 61.6, and 71.6% retention.	The scenario is more optimistic than `S10`, but adoption remains contingent on enterprise readiness, integration cost, and the lag from prototype development to production use.
Aggregation statement	Scenario extracts are not the paper-level prediction.	The final policy report should state that $\hat{y}_p$ is obtained by averaging final-pass estimates across all sampled scenarios at the ToC phase mapped to each target indicator.	Individual scenario rows are included only to make the averaged estimate interpretable and auditable.

Table 13: Compact examples of final-report extracts from randomly sampled scenario logs. The table combines the numeric target estimate, the ToC pathway evidence, and the caveat that should accompany the reported value.

These examples show the intended reporting style. For Policy A, approved projects are useful as a pipeline diagnostic, but the report foregrounds deployed 100E projects and enterprise AI adoption because they correspond to the outcome-facing quantities in the ToC schema. For Policy B, the report keeps graduate employment and international collaboration together while explicitly stating that the two indicators depend on different causal mechanisms. In all cases, the final report keeps the estimate and its caveat adjacent, so that a reader can distinguish policy-grounded constraints from assumptions inferred during simulation.

I Human Evaluation

We conducted a systematic human evaluation on the same randomly sampled 10% subset of Policy C used in Section 5.3. The evaluation covered 60 utterances per method, consisting of three utterances per scenario distributed across the five ToC phases. Five annotators independently evaluated four dimensions: evidence support (0 = unsupported, 1 = partially supported, and 2 = fully supported), hallucination (binary), rule violation (binary), and coherence (a five-point Likert scale). We aggregated hallucination and rule violation by majority vote and report the mean ratings for evidence support and coherence.

Table 14: Human evaluation of reasoning quality on Policy C. Higher evidence-support and coherence scores are better, whereas lower hallucination and rule-violation rates are better.

Method	Evidence (0–2)	Halluc. (%)	Rule viol. (%)	Coherence (1–5)
SOTOPIA-S4	0.9	18.3	8.3	3.4
OASIS	1.0	6.7	11.7	3.9
TIGRIS (*Fwd→Bwd*)	**1.3**	**3.3**	**0.0**	**4.3**

Inter-annotator agreement was substantial for the two binary dimensions, with Fleiss' $\kappa = 0.74$ for hallucination and $\kappa = 0.78$ for rule violation. Krippendorff's α was 0.69 for evidence support and 0.73 for coherence.

J KG Extraction Quality Evaluation

Because TIGRIS uses an automatically extracted policy knowledge graph as the institutional substrate for persona construction and policy-impact simulation, we evaluate the intrinsic quality of the KG extraction process. This section reports two complementary analyses. First, we characterize whether the extracted graphs instantiate the intended IAD/IG 2.0 structure across policy documents. Second, we conduct human validation to assess whether extracted graph elements faithfully represent the source documents.

J.1 Coverage Characterization

Coverage characterization evaluates the structural completeness of the extracted policy KGs before human accuracy scoring. This analysis does not measure correctness; rather, it checks whether the IAD/IG 2.0 ontology is instantiated in a way that can support downstream simulation. In particular, we examine graph scale, IAD node-family coverage, regulative-edge counts, and rule-attribute coverage among extracted regulative edges.

Policy	Chunks	Nodes	Edges	Reg.	IAD Cov.
A	88	218	280	58	5/5
B	42	96	134	37	5/5
C	6	20	59	25	5/5
D	48	110	156	29	5/5
Total	184	444	629	149	–

Table 15: Extracted KG scale and structural coverage. Reg. denotes regulative edges. IAD Cov. indicates whether the five IAD node families used in TIGRIS are instantiated in each policy graph.

Policy	D	C-act	C-ex	O	Src.
A	100.0	46.6	22.4	0.0	99.0
B	100.0	64.9	35.1	5.4	98.5
C	100.0	76.0	48.0	8.0	100.0
D	100.0	55.2	31.0	3.4	98.1
Macro Avg.	100.0	60.7	34.1	4.2	98.9

Table 16: Rule-attribute coverage among extracted regulative edges. D denotes deontic modality, C-act activation condition, C-ex execution constraint, O or-else consequence, and Src. source evidence availability. Values are percentages.

As shown in Table 15, all four policy graphs instantiate the IAD node families required by TIGRIS, suggesting that the ontology can represent diverse AI talent-development policy documents. Table 16 shows a consistent rule-attribute pattern. Deontic modality is explicit in all regulative edges, whereas activation conditions, execution constraints, and or-else consequences are less frequently observed. This pattern is expected in policy documents: eligibility and obligation statements are often explicit, while sanctions, exceptions, and execution constraints are frequently implicit, distributed across clauses, or omitted from the document.

J.2 Human Validation Protocol

We further evaluate extraction accuracy through human annotation. For short documents, we annotated all available source chunks; for longer documents, we sampled up to 40 chunks using a stratified strategy. The sampling covered policy goals, actor roles, eligibility conditions, resource allocation, application or implementation procedures, evaluation criteria, reporting duties, and exceptions or restrictions.

Two annotators with backgrounds in policy analysis independently annotated institutionally relevant spans and mapped them to the TIGRIS ontology. Disagreements were resolved through adjudication to produce the final reference set. Table 17 summarizes the annotation sample.

Annotators identified institutionally relevant spans, assigned IAD node and IG 2.0 edge categories, annotated text-supported ABDICO attributes, and linked each element to its minimal evidence span.

Policy	Ann. Ch.	Gold N.	Gold E.	Crit. R.
A	40	125	164	34
B	32	86	112	29
C	6	24	44	18
D	34	88	116	24
Total	112	323	436	105

Table 17: Human annotation sample for KG extraction validation. Ann. Ch. denotes annotated chunks, Gold N. gold nodes, Gold E. gold edges, and Crit. R. critical rules.

Gold annotation units. The gold annotation contains three types of units. First, annotators identified institutional nodes and assigned IAD-based node categories, including Participants, Positions, Actions, Information, and Costs/Benefits. Second, annotators identified policy relations and assigned IG 2.0-based edge categories, including regulative, constitutive, and structural relations. Third, for regulative edges, annotators marked rule attributes corresponding to the relevant IG 2.0 components, including deontic modality, action or aim, object, activation condition, execution constraint, and or-else consequence when explicitly specified. Each node and edge was linked to the minimal source span supporting the extraction.

Metrics. We report node-level precision, recall, and F1; edge-level precision, recall, and F1; regulative-rule attribute F1; evidence grounding accuracy; and critical-rule recall. A predicted node is counted as correct when its label is semantically equivalent to a gold node and its IAD category matches. A predicted edge is counted as correct when its source node, target node, normalized relation type, and IG 2.0 edge category match a gold edge. Regulative-rule attribute F1 evaluates whether the extracted rule attributes correctly capture the deontic modality, condition, constraint, and consequence information specified in the source document. Evidence grounding accuracy measures whether the extracted graph element is supported by the cited source span.

Critical-rule recall measures whether TIGRIS recovers rules that directly constrain agent reasoning during simulation. We operationalize critical rules as rules satisfying at least one of the following conditions: (i) the rule contains a binding deontic modality such as *must* or *cannot* attached to an actor-type node; (ii) the rule specifies numerical bounds, such as funding limits, cohort caps, support periods, or quota constraints; or (iii) the rule defines eligibility conditions that delimit beneficiary populations.

J.3 Human Validation Results

Policy	Node P	Node R	Node F1	Edge P	Edge R	Edge F1
A	0.85	0.81	0.83	0.76	0.68	0.72
B	0.81	0.77	0.79	0.72	0.64	0.68
C	0.84	0.80	0.82	0.75	0.67	0.71
D	0.80	0.76	0.78	0.71	0.64	0.67
Macro Avg.	0.83	0.79	0.81	0.74	0.66	0.70

Table 18: Human validation of node and edge extraction. Macro averages are unweighted means across policies.

Policy	Rule F1	Evid.	Crit. R	Node κ	Edge κ	Deon. κ
A	0.65	0.89	0.80	0.76	0.68	0.82
B	0.61	0.86	0.76	0.74	0.67	0.80
C	0.64	0.88	0.79	0.78	0.70	0.83
D	0.60	0.85	0.75	0.72	0.65	0.78
Macro Avg.	0.63	0.87	0.78	0.75	0.68	0.81

Table 19: Validation of rule attributes, evidence grounding, critical-rule recall, and inter-annotator agreement. Rule F1 evaluates regulative-rule attributes. Evid. denotes evidence grounding accuracy. Crit. R denotes critical-rule recall. κ values are pre-adjudication Cohen's κ.

Table 18 shows that TIGRIS achieves stronger performance on IAD node extraction than on IG 2.0 edge extraction. The macro-average node F1 is 0.81, indicating that the framework reliably identifies institutionally relevant actors, roles, information objects, actions, and resources. Edge extraction is more difficult, with a macro-average edge F1 of 0.70, because policy relations require correctly identifying not only the two endpoints but also the relation type and the institutional condition under which the relation applies.

Table 19 further shows that regulative-rule attributes remain the most challenging component. The macro-average Rule F1 is 0.63, reflecting the difficulty of extracting activation conditions, execution constraints, and or-else consequences from natural-language policy documents. Evidence grounding accuracy is higher at 0.87, suggesting that most extracted graph elements can be traced to supporting document spans. Critical-rule recall is 0.78, indicating that TIGRIS recovers most simulation-relevant eligibility, obligation, prohibition, funding, and quota rules, while still leaving room for improvement in implicit or exception-heavy policy clauses.

Overall, the validation supports the use of the extracted KG as a structured institutional reference for downstream simulation, while also clarifying the main residual error source: TIGRIS is more reliable at identifying institutional entities and explicit deontic rules than at recovering implicit conditions, execution constraints, and or-else consequences.

K Artifact Use and Computational Details

TIGRIS uses publicly available policy documents, synthetic persona resources, hosted LLM backbones, and baseline simulation systems for research evaluation. The policy documents are used only as pre-implementation inputs for policy knowledge graph construction, while post-implementation indicators are reserved for evaluation. The Nemotron-Personas resources are used to retrieve role-compatible synthetic personas for actor nodes, rather than to identify or model real individuals. OASIS and SOTOPIA-S4 are used as comparative LLM-based simulation baselines under the same parsed policy documents, target indicators, persona pool, and scenario configuration.

All experiments are inference-only. We do not train or fine-tune any model. We use DeepSeek-V4-Flash, GPT-4o-mini, and Gemini-2.0-Flash as hosted LLM backbones. For each policy and model-method configuration, we run $N = 200$ scenarios with $M = 5$ agents per scenario and compare four chain configurations when applicable. We do not perform hyperparameter search because TIGRIS is evaluated as an inference-only simulation framework. The controlled configuration variables are the chain configuration Δ, the number of scenarios N, and the number of agents per scenario M, which are fixed across model-method comparisons unless otherwise stated.