Graph Consistency Rule Mining with LLMs: an Exploratory Study

Rules Generated by LLMs on the property graphs

WWC-2019	**Zero-shot prompt**	**Supp-Cover (%)-Conf (%)**
LLaMA-3	(1) Each person node should have a unique id property	2020 - 100 - 100
	(2) Each person node should have a name and dob property	1678 - 82.69 - 83
	(3) Each match node should have a unique id property	284 - 100 - 100
	(4) Each match node should have a date and stage property	284 - 100 - 100
	(5) Each squad node should have a unique id property	136 - 100 - 100
	(6) Each squad node should have an id property	136 - 100 - 100
	(7) Each team node should have a unique id property	36 - 100 - 100
	(8) Each team node should have a name property	36 - 100 - 100
	(9) A person node can only be connected to a match node through a ‘PLAYED_IN‘ or ‘SCORED_GOAL‘ relationship	1576 - 100 - 100
	(10) A squad node can only be connected to a tournament node through a ‘FOR‘ relationship	136 - 100 - 100
	(11) A team node can only be connected to a squad node through a ‘NAMED‘ relationship	36 - 100 - 100
	(11) A team node can only be connected to a tournament node through a ‘PARTICIPATED_IN‘ relationship	36 - 100 - 100
	(12) A match node can only be connected to a tournament node through an ‘IN_TOURNAMENT‘ relationship	284 - 100 - 100
Mixtral	(1) Each tournament should have a unique identifier	8 - 100 - 100
	(2) Each match should have a unique identifier within its respective tournament	8 - 100 - 100
	(3) Each squad should have a unique identifier within its respective tournament	136 - 100 - 100
	(4) For any given tournament, every team should be associated with a squad	136 - 100 - 100
	(5) A team should be involved in at least one match for each tournament they participate in	136 - 100 - 100
	(6) A match should have a score for both teams (no NULL values) if their match has a score defined	284 - 100 - 100
	(7) Each player should have a unique identifier	2020 - 100 - 100
	(8) A player should be associated with a squad, and that squad should belong to the tournament for which the player has played a match	7684 - 81.1 - 81.1
	(9) A player should score a goal in a match only if they played in that match	903 - 99.2 - 24.5 5
WWC-2019	**Few shot learning**	
LLaMA-3	(1) Ensures that the id_str in Tweet nodes is unique. Each Tweet should have a unique id_str	2020 - 100 - 100
	(2) Validates that the date values in Match nodes fall within a valid range (e.g., from 1900 to 2030)	284 - 100 - 100
	(3) Ensures that the stage values in Match nodes are consistent and contain no invalid values or errors	284 - 100 - 100
	(4) Ensures that Team nodes have valid and non-null name attributes	36 - 100 - 100
	(5) Ensures that each Match node is associated with exactly two Team nodes	284 - 100 - 100
	(6) Ensures that squad names are unique within a tournament, meaning no duplicate squad names for a given tournament	136 - 100 - 100
	(7) Ensures each Person node is associated with at most one Team node	1928 - 95.3 - 100
	(8) Ensures that only Person nodes associated with a Team participating in a Match can score a goal	756 - 42.5 - 100
Mixtral	(1) Each Match node must have a unique ID	2020 - 100 - 100
	(2) Each Match node must have a unique date and phase in a specific tournament	39 - 100 - 100
	(3) Each Team node must have a unique name in a specific tournament	36 - 100 - 100
	(4) Each Tournament node must have a unique name	8 - 100 - 100
	(5) Each Squad node must have a unique ID	136 - 100 - 100
	(6) Ensure each Person-Squad relationship has exactly one Squad representing the Person	1268 - 100 - 66
	(7) Ensure that each Match-Team relationship within a specific tournament is valid	568 - 100 - 100
	(8) Ensure no duplicate SCORED_GOAL relationships exist with the same scoring minute	907 - 42.5 - 100

Table 1: Rules Generated by LLMs on the WWC-2019 graph with Sliding Window Attention

WWC-2019	Zero-shot prompt	Supp-Cover (%)-Conf (%)
LLaMA-3	(1) Each node should have a unique ID within its respective category (e.g., player, team, coach)	2468 - 100 - 100
	(2) Each node should have a consistent label corresponding to its category (e.g., player, team, coach)	2468 - 100 - 100
	(3) A coach can coach only one team at a time	77 - 100 - 75
	(4) A player can only belong to one squad at a time	1268 - 100 - 66
	(5) The IN_SQUAD relationship should only exist between a player and a team	2880 - 100 - 100
	(6) The COACH_FOR relationship should only exist between a coach and a team	140 - 100 - 100
	(7) The REPRESENTS relationship should only exist between a player/coach and a team	1929 - 100 - 100
Mixtral	(1) Each player should be uniquely represented in the graph. There should not be two nodes with the same id under the REPRESENTS relationship	2020 - 100 - 100
	(2) A player cannot be in two different squads at the same time. There should not be two IN_SQUAD relationships for the same player and different squads	1268 - 100 - 66
	(3) A player cannot be in the same squad more than once. There should not be multiple IN_SQUAD relationships between the same player and the same squad	1268 - 100 - 66
	(4) If a player is substituted in a game, the minuteOn value in the PLAYED_IN relationship should be less than the minute value in the SCORED_GOAL relationship (if it exists)	0 - 0 - 0
	(5) A player cannot score two goals in the same minute of the same game. There should not be two SCORED_GOAL relationships for the same player, same game, and same minute	907 - 42.5 - 100
	(6) A coach cannot coach two different teams at the same time. There should not be two COACH_FOR relationships for the same coach and different teams	77 - 100 - 75
WWC-2019	**Few shot learning**	
LLaMA-3	(1) Each REPRESENTS relationship should have a unique id	2468 - 100 - 100
	(2) A player can only be part of one squad at a time	1268 - 100 - 66
	(3) A player can only play in one match at a time	1565 - 100 - 100
	(4) A coach can only coach one team at a time	77 - 100 - 75
	(5) A player can only score one goal per match	636 - 42.5 - 100
	(6) A player can only be subbed on or off once per match	0 - 0 0
Mixtral	(1) Each relationship should have unique nodes at its ends. This means that a relationship cannot connect the same two nodes more than once, regardless of the relationship type	14528 - 100 - 100
	(2) The relationship types should match the node types they are connecting. For example, a ‘PLAYED_IN‘ relationship should only connect a Person node to a Match node	7684 - 100 - 100
	(3) Each Player node should have only one ‘REPRESENTS‘ relationship to a Person node. This ensures that each player is represented by only one person	1928 - 95.3 - 100
	(4) For ‘PLAYED_IN‘ relationships with a ‘minuteOn‘ property, the sum of ‘minuteOn‘ for a player in a match should not exceed the total match minutes	0 - 0 - 0
	(5) For ‘PLAYED_IN‘ relationships, if a ‘type‘ property is present, it should be consistent for all relationships of a player in a match	7684 - 20.3 - 100

Table 2: Rules Generated by LLMs on the WWC-2019 graph with RAG

Cybersecurity	Zero-shot prompt	Supp-Cover (%)-Conf (%)
LLaMA-3	(1) Each node should have a unique neo4jImportId	953 - 1.0 - 1.0
	(2) All nodes with a domain property should have a valid domain name	953 - 100 - 100
	(3) The owned property should only be True or False	603 - 63 - 100
	(4) The enabled property should only be True or False	600 - 63 - 100
	(5) User nodes should have a valid displayname and username	300 - 100 - 100
	(7) Computer nodes should have a valid operatingsystem and domain	301 - 100 - 100
	(7) Group nodes should have a valid displayname and domain	308 - 100 - 100
	(8) OU nodes should have a valid displayname and domain	21 - 100 - 100
	(9) GPO nodes should have a valid displayname and domain	22 - 100 - 100
	(10) Relationships should only connect nodes of compatible types (e.g., User-Computer, Group-User, etc.)	0 - 0 - 0
Mixtral	(1) If a node has the label Computer, it must have the properties operatingsystem, name, objectid, and enabled	300 - 97.2 - 99.67
	(2) If a node has the label User, it must have the properties displayname, pwdlastset, domain, lastlogon, name, objectid, and enabled	299 - 99.67 - 99.76
	(3) If a relationship has the type HAS_SESSION, it must involve a Computer node and a User node	264 - 100 - 100
	(4) If a relationship has the type MEMBER_OF, it must connect a User node to a Group or Domain node	1043 - 76.7 - 76.6
	(5) If a relationship has the type ADMIN_TO, it must connect a User or Group node to a Domain or Computer node	Support: 1347 - 100 - 100
	(6) For each Domain node, there must be at most one Domain node connected via the PARENT_OF relationship	0 - 0 - 0
	(7) For each Domain node, there must be at least one Domain node connected via the SUB_DOMAINS relationship	0 - 0 - 0
	(8) For each Group node, there must be at least one User node connected via the MEMBER_OF relationship	262 - 100 - 100
	(9) For each Domain node, there must be at least one OU node connected via the CONTAINS relationship	1 - 0 0
	(10) For each GPO node, there must be exactly one OU node connected via the AFFECTS relationship	0 - 0 - 0
	Few shot learning	
LLaMA-3	(1) Each node should have a unique neo4jImportId property	953 - 100 - 100
	(2) Each node should have a unique objectid property	948 - 99.5 - 99.8
	(3) Only allowed labels: User, Group, Domain, OU, GPO, and Computer	953 - 100 - 100
	(4) The owned property should be a boolean value	603 - 63 - 100
	(5) The domain property should be a string value matching domain format	953 - 100 - 100
	(6) The operatingsystem property should match valid OS names	300 - 100 - 100
	(7) The name property should match username or computer name format	953 - 100 - 100
	(8) The enabled property should be a boolean value	600 - 63 - 100
	(9) Only allowed relationship types: HAS_SESSION, MEMBER_OF, CONTAINS, GP_LINK, etc	3762 - 80 - 81
Mixtral	(1) Each node in the database should have a unique neo4jImportId value	953 - 100 - 100
	(2) Each node in the database should have a unique objectid value	948 - 99.5 - 99.8
	(3) Only the allowed node labels (User, Group, Domain, OU, GPO, and Computer) should be present	953 - 100 - 100
	(4) A Computer node should not have a MEMBER_OF relationship with a Domain node	301 - 100- 100
	(5) A GPO node should be linked to either a Domain node or an OU node through a GP_LINK relationship	21 - 60 - 60

Table 3: Rules Generated by LLMs on the Cybersecurity graph with Sliding window attention

Cybersecurity	Zero-shot prompt	Supp-Cov (%)-Conf (%)
LLaMA-3	(1) Each node should have at most one unique neo4jImportId	953 - 100 - 100
	(2) If node A has a MEMBER_OF relationship with node B, then node B should also have a MEMBER_OF relationship with node A	2 - 2/1361 - 2/1361
	(3) If node A has an ADMIN_TO relationship with node B, and B has an ADMIN_TO relationship with node C, then node A should also have an ADMIN_TO relationship with node C	0 - 0 - 0
	(4) Each node should have at most one HAS_SESSION relationship with another node	114 - 43 - 100
	(5) Each node should have at most one CAN_RDP relationship with another node	12 - 12/60 - 12/60
	(6) Each node should have at most one EXECUTE_DCOM relationship with another node	19 - 19/60 - 19/60
	(7) Each node should have at most one ALLOWED_TO_DELEGATE relationship with another node	14 - 30 - 0.67
Mixtral	(1) Each node should have only one neo4jImportId property	953 - 100 - 100
	(2) A MEMBER_OF node should not have more than one outgoing relationship. This ensures that a node is not a member of multiple entities	340 - 100 - 59
	(3) A node connected to the GENERIC_ALL node should not have any other relationships, ensuring that it only inherits properties from the GENERIC_ALL node	1 - 50 - 50
	(4) The enabled property should only be True or False	600 - 63 - 100
	(5) A node with the CONTAINS label should not have any incoming relationships, ensuring it does not inherit properties from other nodes	0 - 0 - 0
	(6) A node with the MEMBER_OF label should not have any incoming relationships, ensuring it does not inherit properties from other nodes	0 - 0 - 0
Cybersecurity	**Few shot learning**	
LLaMA-3	(1) Ensure that each node has a unique id property, which helps in distinguishing nodes in the graph	953 - 100 - 100
	(2) Ensure that each node has a label from the allowed set: User, Group, Domain, OU, GPO, or Computer, ensuring proper classification of nodes	953 - 100 - 100
	(3) Ensure that MEMBER_OF relationships point from a User or Group node to a Group node	1043 - 77 - 98
	(4) Ensure that ADMIN_TO relationships point from a Group node to a Computer node, indicating correct access delegation	1347 - 100 - 100
	(5) Ensure that HAS_SESSION relationships point from a User node to a Computer node, indicating user sessions on computers	264 - 100 - 100
	(6) Ensure that CAN_RDP relationships point from a User node to a Computer node, indicating remote desktop permissions	30 - 50 - 100
	(7) Ensure that ALLOWED_TO_DELEGATE relationships point from a User node to a Computer node, indicating delegation rights for remote desktop access	30 - 100 - 100
Mixtral	(1) Ensure that each relationship has a unique relationship Type property. This helps distinguish and identify each relationship	4698 - 100 - 100
	(2) Ensure that the isacl property for GENERIC_ALL nodes is consistently set to True, indicating they are ACL entries	901 - 100 - 100
	(3) Ensure that the label property for all nodes is one of User, Group, Domain, OU, GPO, Computer, or GENERIC_ALL	953 - 100 - 100
	(4) Ensure that no nodes are isolated and all nodes are connected within the graph structure	952 - 99.8 - 100
	(5) The highvalue property should be consistent across nodes of the same type	1 - 100 - 100

Table 4: Rules Generated by LLMs on the Cybersecurity graph with RAG

Twitter	**Zero-shot prompt**	**Support-Coverage (%)-Confidence (%)**
LLaMA-3	(1) Each tweet node should have a unique *id* property	2407 - 100 - 100
	(2) All URL properties in Link nodes should be valid URLs (starting with 'http://' or 'https://')	1561 - 100 - 100
	(3) All User nodes should have a non-empty name property (screen_name)	38986 - 100 - 100
	(4) All Hashtag nodes should have a non-empty name property	344 - 100 - 100
	(5) Each Tweet node should have at most one Source node connected to it through the USING relationship	2146 - 89.15 - 100
	(6) Ensure that the relationships between nodes are valid and consistent (e.g., Tweet node connected to User node through POSTS)	2146 - 89 - 100
	(7) Each User node should have a unique id property	0 - 0 - 0
	(8) Ensure that FOLLOWS relationships between User nodes are valid and consistent (i.e., no cycles, only User to User)	44630 - 100 - 100
Mixtral	(1) A User node must have a unique screen_name property	38986 - 100 - 100
	(2) A Tweet node should not have more than one created_at property	2407 - 100 - 100
	(3) A User node should not be connected to a Tweet node by both MENTIONS and MENTIONS, Me relationships	3459 - 100 - 100
	(4) A User node linked with the FOLLOWS relationship should have a name property	34507 - 100 - 100
	(5) A Hashtag node should not be connected to a Tweet node by both TAGS and TAGS, Me relationships	1439 - 100 - 100
	(6) A A Tweet node should have a text property with a maximum length of 280 characters.	2146 - 100 - 100
	(7) A Tweet node should have a unique id property	2407 - 100 - 100
	(8) A Link node should have a unique url property	1561 - 100 - 100
	(9) A Tweet node should have zero or more RETWEETS relationships	286 - 11.89 - 11.89
	(10) A Tweet node should have zero or one created_by relationship	2407 - 100 - 100
Twitter	**Few shot learning**	
LLaMA-3	(1) Each user node should have a unique screen_name property to ensure that no duplicate users exist	38986 - 100 - 100
	(2) The followers property should be a non-negative integer	38690 - 100 - 100
	(3) The following property should be a non-negative integer	38690 - 100 - 100
	(4) The profile_image_url property should be a valid URL	38960 - 100 - 100
	(5) The location property should follow a consistent format (e.g., city, state, country)	32 - 0.082 - 0.082
	(6) The url property should be a valid URL	16997 - 100 - 100
	(7) The statuses property should be a non-negative integer	4052 - 100 - 100
Mixtral	(1) Ensure that every tweet has a unique id_str property	2146 - 100 - 100
	(2) Ensure that no two users have the same name and screen_name combination	38986 - 100 - 100
	(3) Ensure that Hashtag nodes are only connected to Tweet nodes	1493 - 100 - 100
	(4) Ensure that relationships going out from User nodes have the label "FOLLOWS" and connect to a User node	44630 - 100 - 100
	(5) Ensure that a user does not have multiple nodes with the same location as their home_location	7808 - 79.33 - 79.33
	(6) Ensure that User, Me nodes only have Tweet and Follows relationships	2146 - 4.59 - 4.59
	(7) Ensure that a tweet does not have multiple nodes with different labels and urls	129 - 67.64 - 79.24

Table 5: Rules Generated by LLMs on the Twitter graph with Sliding window attention

Twitter	**Zero-shot prompt**	**Support-Coverage (%)-Confidence (%)**
LLaMA-3	(1) A Tweet node can only be connected to one User node with the POSTS relationship type	2147 - 89 - 100
	(2) A Tweet node can only be connected to the same User node (Me) with the MENTIONS relationship type once	899 - 37 - 30
	(3) A Tweet node can only be connected to one other Tweet node with the REPLY$_{T}Orelationshiptype$	504 - 100 - 100
	(4) A Tweet node can only be connected to one other Tweet node with the RETWEETS relationship type	268 - 100 - 100
	(5) A User node can only be connected to the same User node (Me) with the FOLLOWS relationship type once	0 - 0 - 0
	(6) A Tweet node can only be connected to one Link node with the CONTAINS relationship type	1499 - 89 - 100
	(7)A Tweet node can only be connected to one Hashtag node with the TAGS relationship type	389 - 47 - 100
	(8) A Tweet node can only be connected to one Source node with the USING relationship type	2146 - 100 - 100
Mixtral	(1) In the graph, a Tweet can only have a REPLY_TO or RETWEETS relationship with another Tweet. It cannot have both or neither	722 - 32 - 32
	(2) n the graph, a User can only have a FOLLOWS relationship with another User. It cannot follow a Tweet directly	44630 - 100 - 100
	(3) In the graph, a User or User, Me can only have a POSTS relationship with a Tweet. It cannot post to another User directly	2146 - 100 - 100
	(4) In the graph, a Tweet can only have a TAGS relationship with a Hashtag. It cannot be tagged with a User or any other node	389 - 47 - 100
	(5) In the graph, a Tweet can only have a CONTAINS relationship with a Link. It cannot contain a User or any other node	1499 - 92 - 100
	(6) In the graph, a Tweet can only have a USING relationship with a Source. It cannot use a User or any other node	2146 - 100 - 100
	(7) In the graph, a User, Me can only have a FOLLOWS relationship with a User. It cannot follow a Tweet directly	0 - 0 - 0
Twitter	**Few shot learning**	
LLaMA-3	(1) Each user node should have a unique screen_name property to ensure that no duplicate users exist	38986 - 100 - 100
	(2) Each Tweet node should have a valid created_at timestamp in the format yyyy-MM-dd'T'HH:mm:ss'Z' to ensure consistent date formatting	2407 - 100 - 100
	(3) Every Tweet node should have a valid text property to ensure that tweet content is not empty	2416 - 89.16 - 100
	(4) Each Hashtag node should have a unique name property to prevent duplicate hashtags	344 - 100 - 100
	(5) Each Link node should have a valid url property to ensure that links are properly formatted	1561 - 100 - 100
	(6) Every User node should have a valid id property to ensure that user IDs are consistent	0 - 0 - 0
	(7) Each Tweet node should have at most one RETWEETS relationship to another Tweet node to prevent multiple retweet relationships	286 - 100 - 100
	(8) Each User node should have at most one FOLLOWS relationship to another User node to prevent multiple follow relationships	34506 - 88.51 - 100
	(9) Each Tweet node should have at most one MENTIONS relationship to another User node to prevent multiple mention relationships	718 - 66.40 - 100
Mixtral	(1) Each User node should have a unique screen_name property to ensure no duplicate user accounts exist	38986
	(2) Nodes labeled Tweet should have a created_at property to ensure the timeline of tweets is correct	2407 - 100 - 100
	(3) Nodes labeled Hashtag should have a unique name property to ensure no duplicate hashtags exist	344 - 100 - 100
	(4) The MENTIONS relationship should only exist between User and Tweet nodes to maintain the integrity of the graph	3459 - 100 - 100
	(5) The REPLY_TO relationship should only exist between Tweet nodes to maintain the integrity of the graph	504 - 100 - 100
	(6) The RETWEETS relationship should only exist between Tweet nodes to maintain the integrity of the graph	268 - 100 - 100
	(7) The USING relationship should only exist between Tweet and Source nodes to maintain the integrity of the graph	2146 - 100 - 100
	(8) The FOLLOWS relationship should only exist between User nodes to maintain the integrity of the graph	44630 - 100 - 100

Table 6: Rules Generated by LLMs on the Twitter graph with RAG