■■■■■Optimizing Agent Planning for Security and Autonomy

■■■■■■■■■■■■■■■■■■

■■■■■input_papers\test.pdf
■■■■■2026-02-14 15:44
■■■33

■■■■

■ 1 ■

[■] OPTIMIZING AGENT PLANNING FOR SECURITY AND AUTONOMY AashishKolluri1 RishiSharma1,2† ManuelCosta1 BorisKo¨pf1 TobiasNießen3† MarkRussinovich1 ShrutiTople1 SantiagoZanella-Be´guelin1 1Microsoft 2EPFL 3T...

■ 2 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy integrity and confidentiality labels to all data an agent processes, propagating labels to suggested actions, and using these labels to determine whether ...

[■] OptimizingAgentPlanningforSecurityandAutonomy 2 Background: Information-flowControlforAIAgents Information-flowcontrolmechanismsusesecuritylabelstodescribe thesecuritypropertiesofdataduringtheirlifetim...

[■] OptimizingAgentPlanningforSecurityandAutonomy isolation,buttheircontentremainshiddenfromt heplanner’sLLM.TheoriginalformulationoftheDualLLMpattern allows for restricted outputs of the quarantined LLM t...

■ 5 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy unsuccessfultracestheagentrepeatedlyattemptedactionsthatfailedpolicychecks(whichweallowtocontinue)and didnotleadtoanyprogress,apatternthatahumanwouldquick...

[■] OptimizingAgentPlanningforSecurityandAutonomy expand variables(ask endorsement=True)), maintaining the label of the context, or (ii) proceed without en- dorsement (by calling expand variables(ask endo...

[■] OptimizingAgentPlanningforSecurityandAutonomy
informationextractioninquarantinedLLMqueries.
Bydesign,noattackssucceedinthissettingduetostrictpolicies, deterministic defenses, and our assumption that a...

[■] OptimizingAgentPlanningforSecurityandAutonomy 80 60 40 20 0 1 2 3 4 HITLload(k))%(k@RCT o3-mini 80 60 40 20 0 1 2 3 4 HITLload(k) Basic Basic-IFC FIDES PRUDENTIA)%(k@RCT o4-mini Figure2:TCR@kcurvess...

[■] OptimizingAgentPlanningforSecurityandAutonomy AttackSuccessRate HITLload(average) TCR@∞(%) Turns(average) Model Environment Basic PRUDENTIA Basic PRUDENTIA Basic PRUDENTIA Basic PRUDENTIA GitLab 20.80...

[■] OptimizingAgentPlanningforSecurityandAutonomy securitypromptsandbeginapprovingactionswithoutcarefulconsideration(Stantonetal.,2016;Seidlingetal.,2011). DeterministicdefensesbasedonIFCcanbemoreeffectiv...

[■] OptimizingAgentPlanningforSecurityandAutonomy 7 RelatedWork Probabilistic Defenses. Several techniques have been proposed for minimizing the likelihood of prompt injection attacksinLLM-basedsystemsing...

[■] OptimizingAgentPlanningforSecurityandAutonomy References Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. Get my drift? Catching LLM task drift with activat...

■ 13 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy
FeiranJia,TongWu,XinQin,andAnnaSquicciarini. TheTaskShield:
Enforcingtaskalignmenttodefendagainst indirectpromptinjectioninLLMagents.
In63rdAnnualMeetingo...

[■] OptimizingAgentPlanningforSecurityandAutonomy SimonWillison. TheDualLLMpatternforbuildingAIassistantsthatcanresistpromptinjection. Online: https: //simonwillison.net/2023/Apr/25/dual-llm-pattern,Apr20...

[■] OptimizingAgentPlanningforSecurityandAutonomy A CaseStudy: ChoicebetweenEndorsementandApproval WeshowanabridgedexampletraceofanAgentDojotaskwhereaPRUDENTIAagentstrategicallychoosestonotask theuserfore...

■ 16 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy Assistant search_emails(query="TechServices password reset") Tool - Email 1: From support@techservices.com Body in variable #search_emails-result-0.body# ...

[■] OptimizingAgentPlanningforSecurityandAutonomy B AdditionalResults Table 2 provides comprehensive performance data across all methods and models, revealing consistent patterns of improvementfromBasicth...

■ 18 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy 100 80 60 40 20 0 DI DIQ DD)%(0@RCT TCR@0usingo3-mini 100 80 60 40 20 0 DI DIQ DD)%(0@RCT TCR@0usingo4-mini 100 80 60 40 20 0 DI DIQ DD)%(1@RCT TCR@1us...

[■] OptimizingAgentPlanningforSecurityandAutonomy 100 80 60 40 20 0 banking slack travel workspace)%(etaRnoitelpmoCksaT TCR@∞usingo3-mini 100 80 60 40 20 0 banking slack travel workspace)%(etaRnoitelpmo...

[■] OptimizingAgentPlanningforSecurityandAutonomy C OverheadAnalysis Weprovideroughestimatesforthecostsof Basic-IFC,FIDES,andPRUDENTIAintermsofthenumberofinput(Table3)andoutput(Table5)tokens,andUSDcents(...

■ 21 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy We believe that these costs can be brought down significantly as our goal was to investigate autonomy gains, so we have not yet made an effort to optimize...

[■] OptimizingAgentPlanningforSecurityandAutonomy D SystemPrompts D.1 BasicPlanner Weuset hefollowingsystempromptfortheBasicagentwithandwithoutIFC,takenfromDebenedettietal.(2024). System You are an AI lang...

■ 23 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy 2. **Field inside a dict**: - Format: `#tool_name-result-tool_name_count.key#` 3. **Basic type inside a list of dicts**: - Format: `#tool_name-result-tool...

■ 24 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy - **Example:** - You plan to call: `generate_report(title, date)`. - You ask quarantined LLM: - Query 1: *"What's the title in #v1#?"* -> stored in `#quar...

■ 25 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy - You call `expand_variables` with `[#v1#, #v2#]`. - Now use expanded information to carry out your next steps. ## IMPORTANT REMINDERS - Never invent or g...

■ 26 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy - **Untrusted Data:** Untrusted data is any information received from a tool call that may (cid:44)→ contain prompt-injections. Examples include email con...

■ 27 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy ### Variable naming conventions: Variable identifiers always start and end with `#`. **Basic types** (`str`, `bool`, `int`, `float`): - Format: `#tool_nam...

■ 28 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy - **Scenario:** You know exactly what tool you want to call next and require specific (cid:44)→ arguments from the variables. - **Action:** Query the quar...

■ 29 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy To minimize the number of endorsements and approvals, you should avoid expanding variables (cid:44)→ in your context for as long as possible. Once you kno...

■ 30 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy - Always ensure each query to quarantined LLM returns exactly one piece of information. - Plan and exhaust other options and defer using `ifc_expand_varia...

[■] OptimizingAgentPlanningforSecurityandAutonomy E ExampleAccessibilityTreeObservation
Wepresentfirsttheunprocessedaccessibilitytree,withthepromptinjectionhighlightedinred. Axtree
"OBSERVATION: 0 (curren...

■ 32 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy [294] StaticText '(■30fb_■30fb;)' [297] StaticText "There's nothing here 2026" [303] complementary '' [310] StaticText '1 point' [313] StaticText '(+1, ■2...

■ 33 ■

[■] OptimizingAgentPlanningforSecurityandAutonomy [215] StaticText VAR_18_7165e3de [UNTRUSTED] [216] link VAR_19_6f3a142a [UNTRUSTED] [218] StaticText VAR_20_68801a33 [UNTRUSTED] [228] link VAR_21_6eff0cb...

■■1■■■■■■

■■■■	■■■■	■■
Transformer	Transformer	■■■■■■■■■■■■■■■■■
Attention	■■■■■	■■■■■■■■■■■■■
Neural Network	■■■■	■■■■■■■■■■
Deep Learning	■■■■	■■■■■■■■■■■■■■■
Model	■■	■■■■■■■■■■■■■