

# Интеллектуальная надстройка управления на основе языковой модели в системах оптимизации в реальном времени

А.А. Карманов

**Аннотация**— В работе представлена концептуальная архитектура включения больших языковых моделей (LLM) в иерархические системы управления типа Real-Time Optimization – Model Predictive Control (RTO–MPC). Предложен надзорный когнитивный уровень, размещённый между экономическим контуром оптимизации и предиктивным регулятором, который обеспечивает контекстно-зависимую адаптацию целевых уставок, ограничений и весовых коэффициентов функционала управления. Взаимодействие между уровнями реализуется через формализованный протокол обмена данными и интерфейс Model Context Protocol (MCP), обеспечивающий доступ к гетерогенным источникам информации — архивным данным, прогнозам, расписаниям и эксплуатационным журналам. Агент на базе LLM интерпретирует эти данные, формирует машинно читаемые рекомендации и передаёт их в слой RTO для проверки достижимости и устойчивости перед применением в MPC. Таким образом, когнитивный модуль интегрируется в систему без нарушения её детерминированности и свойств устойчивости. Предложенный подход сохраняет математическую строгость классической архитектуры RTO–MPC и расширяет её за счёт возможности семантической адаптации к внешнему контексту. Архитектура открывает перспективы создания гибридных когнитивно-инженерных систем управления для холодильных установок, энергетических комплексов и интеллектуальных зданий.

**Ключевые слова**— RTO–MPC, LLM-агент, иерархическое управление, когнитивная оптимизация, семантическая адаптация, интерфейс MCP, предиктивное регулирование.

## I. ВВЕДЕНИЕ

В последние два десятилетия иерархические схемы «оптимизация в реальном времени — предиктивное управление» стали де-факто стандартом для эксплуатации энергоёмких объектов, где экономические решения верхнего уровня должны согласованно исполняться нижним уровнем динамического управления [1]. Разделение временных масштабов позволяет верхнему уровню периодически решать стационарную экономическую задачу с учётом ограничений и тарифных сигналов, тогда как управление на основе прогноза (MPC) на нижнем уровне обеспечивает быстрое действие слежения за целями

при наличии возмущений и неточностей модели [2]. На примерах крупных зданий и кампусов, холодильных станций и районных систем охлаждения показано, что такая декомпозиция даёт измеримую экономию энергии при сохранении требуемого качества регулирования и вычислимой реализации в реальном времени [3]. Одновременно накопленный опыт выявляет задачи согласования слоев, включающие достижимость передаваемых уставок без резких переходов, смещение экономического оптимума из-за несовпадения модели и объекта, а также деградацию производительности при частых внешних воздействиях, когда статические цели уступают траекторным предписаниям. Для смягчения указанных эффектов предложены фильтрация целей и калькуляторы достижимых уставок, модификационная адаптация с корректировкой целевой функции и ограничений по измеренным смещениям и градиентам, а также динамическая оптимизация в реальном времени (RTO) и экономический вариант MPC, интегрирующий экономический критерий в предиктивный регулятор [4]. В холодильных системах ледовых арен добавляются специфические факторы, среди которых выраженная событийность технологического цикла, высокая чувствительность к нарушениям температурных ограничений и неполная информированность о будущих возмущениях [5]. Классическая двухуровневая архитектура выигрывает от учета краткосрочного прогноза возмущений на горизонте MPC, однако практические процедуры интеграции такого прогноза и оперативной перенастройки параметров регулятора остаются трудоемкими и слабо формализованными, особенно при необходимости учитывать внешние источники знаний, включая погодные сводки, расписания мероприятий и эксплуатационные регламенты. В современной литературе все шире рассматриваются надзорные компоненты машинного обучения, способные адаптировать веса, ограничения и уставки при сохранении детерминированной природы нижнего контура [6]. На этом фоне возникает содержательный исследовательский вопрос о потенциале агента на базе большой языковой модели, который, действуя поверх MPC, повышает робастность и экономичность управления за счет контекстно зависимой реконфигурации целей и ограничений без вмешательства

Статья получена 21 июня 2026.

А. А. Карманов, Финансовый университет при Правительстве РФ,  
(email: aleksandr.karmanov@outlook.com)

в алгоритм решения квадратичных задач [7].

Предлагаемая работа анализирует алгоритм двухуровневой схемы, в которой низовой контур сохраняет классическую постановку MPC с явными ограничениями и стандартной укладкой в квадратичную задачу, а над ним функционирует агент, периодически формирующий конфигурацию регулятора на основе текущего состояния, прогнозов событий и внешних знаний. Все решения агента проходят машинно читаемую валидацию по схеме и офлайн проверку достижимости на предсказательной модели до применения, что согласуется с принципами устойчивости и безопасности, разработанными в области RTO–MPC. Формируется практическая процедура включения внешнего контекста в контур управления, которая опирается на проверенную иерархическую парадигму и сохраняет инженерные гарантии предиктивного регулирования.

## II. ПОСТАНОВКА ЗАДАЧИ И МЕТОД

В задачах управления энергоёмкими технологическими системами, включая холодильные контуры ледовых арен, широко применяется иерархическая схема Real-Time Optimization – Model Predictive Control (RTO–MPC) [1]. Она сочетает экономическую оптимизацию на верхнем уровне с динамическим предиктивным управлением на нижнем. Верхний уровень решает задачу минимизации энергетических или стоимостных затрат при соблюдении технологических ограничений, а нижний обеспечивает слежение за целевыми параметрами с учётом динамики объекта и внешних возмущений. Внешние воздействия включают события заливки и иные источники теплопритоков, для которых допускается внешний прогноз. Предсказательная модель имеет вид

$$x_{t+1} = A x_t + B u_t + E \hat{d}_t, \quad y_t = C x_t$$

Где  $x_t$  — состояние,  $u_t \in [0,1]$  — управление,  $y_t$  — температура льда,  $\hat{d}_t$  — прогноз события/возмущения

Цель предиктивного регулятора состоит в слежении за уставкой при ограниченной вариабельности исполнительных воздействий. На горизонте (H) минимизируется квадратичный функционал по приращениям управления  $\Delta u_t$

$$\min_{\Delta u_{t:t+H-1}} \sum_{k=1}^H \lambda_y |y_{t+k|t} - r_{t+k}|^2 + \sum_{k=0}^{H-1} \lambda_u |\Delta u_{t+k}|^2$$

Ограничения вводятся по выходу, управляющему воздействию и его темпу изменения. Для повышения выполнимости допускается использование мягких допусков с дополнительным штрафом в критерии, однако базовая форма может быть записана без смягчений

$$\begin{aligned} y_{\min} &\leq y_{t+k|t} \leq y_{\max}, & u_{\min} &\leq u_{t+k} \\ &\leq u_{\max}, & |\Delta u_{t+k}| &\leq \delta_u, & k \\ &= 0, \dots, H-1 \end{aligned}$$

Где  $r_t$  — уставка,  $H$  — горизонт,  $\lambda_y, \lambda_u > 0$  — веса,  $y_{\min}, y_{\max}, u_{\min}, u_{\max}, \delta_u$  — технологические ограничения. Стандартная укладка по горизонту переводит задачу в квадратичную форму с линейными ограничениями, что обеспечивает решение в реальном

времени малогабаритным QR-решателем при фиксированном горизонте [8]. Однако практика применения показывает, что между уровнями RTO и MPC часто возникает контекстный разрыв: внешние факторы — расписания, погодные условия, регламенты эксплуатации, операционные замечания — не попадают в контур оптимизации, что снижает адаптивность и робастность системы [7].

Для устранения этого ограничения вводится надзорный слой в виде агента на базе большой языковой модели (LLM) [9]. Агент располагается между уровнями RTO и MPC и выполняет когнитивную функцию интерпретации внешнего контекста. Он получает данные из различных источников (временные ряды, прогнозы, расписания, MCP-интерфейсы), анализирует их и формирует структурированные рекомендации по корректировке уставок, весов и ограничений [10].

## III. АРХИТЕКТУРА И НАДЗОРНЫЙ УРОВЕНЬ

Иерархическая структура системы управления (рис.1) основана на концепции многоуровневого согласования решений, при котором каждый уровень функционирует в собственном временном масштабе и обладает различной степенью детализации модели. Верхний уровень RTO (Real-Time Optimization) решает задачу экономической оптимизации, определяя целевые значения технологических параметров и допустимые диапазоны переменных. Этот уровень оперирует усреднёнными характеристиками процесса и обновляется относительно редко — с шагом от нескольких минут до часа.

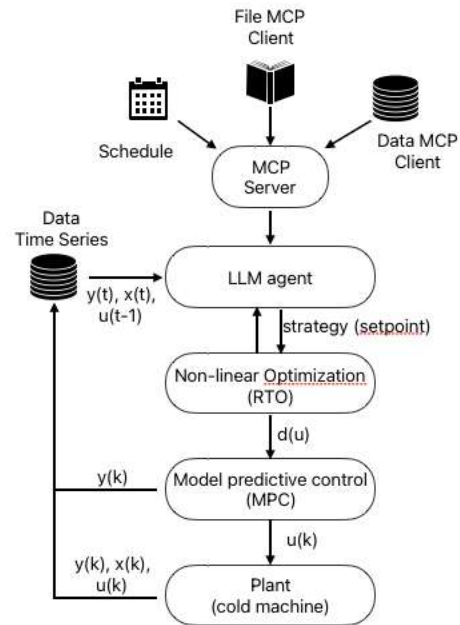


Рис. 1: Архитектура иерархической системы RTO–MPC с надзорным LLM-агентом

Нижний уровень реализует классический MPC, обеспечивающий динамическое слежение за переданными целями при наличии возмущений и ограничений. Предиктивный регулятор формирует оптимальное управляющее воздействие, минимизируя отклонение от уставки при ограниченной

вариабельности исполнительных сигналов. Его временной масштаб соответствует частоте технологического процесса.

Между ними введён надзорный слой, представленный интеллектуальным агентом на базе большой языковой модели (LLM) [4]. Его функция заключается в контекстно-зависимой реконфигурации регулятора, включающей адаптацию весов в целевом функционале, модификацию допустимых диапазонов и формирование новых уставок в зависимости от внешних факторов и прогнозных данных. Агент получает информацию как от RTO (целевые экономические приоритеты), так и из внешних источников — прогнозов погоды, расписаний мероприятий, эксплуатационных регламентов, операторских комментариев.

Поток данных между уровнями организован следующим образом. RTO передаёт LLM-агенту обновлённые цели и экономические ограничения; агент, в свою очередь, формирует структурированный пакет параметров для MPC. Прежде чем параметры применяются, они проходят машинно читаемую проверку достижимости на предсказательной модели — процедура, аналогичная офлайн-валидации, используемой в литературе по безопасному RTO-MPC [1]. Если новая конфигурация удовлетворяет критериям устойчивости и совместимости, она передаётся в MPC; в противном случае активируется резервный профиль управления с последующим уведомлением оператора.

Таким образом, LLM-агент выполняет роль когнитивного посредника между экономическим и динамическим уровнями управления, обеспечивая перевод контекстных знаний в инженерные параметры регулятора при сохранении детерминированной природы нижнего контура. Предлагаемая архитектура объединяет принципы иерархического управления с возможностями адаптивного машинного рассуждения, создавая основу для гибридных систем «RTO-MPC-LLM», способных к самоадаптации без нарушения формальных гарантий выполнимости.

#### IV. LLM АГЕНТ

В предлагаемой архитектуре надзорный агент на базе большой языковой модели (LLM) выполняет функцию когнитивного посредника между экономическим уровнем RTO и уровнем предиктивного управления MPC [4]. Его задача заключается в том, чтобы интерпретировать разнородные источники контекстных данных и преобразовывать их в параметрические рекомендации, согласованные с математической формой задач RTO и MPC. Агент не заменяет существующие алгоритмы оптимизации, а дополняет их слоем рассуждения, который расширяет доступное системе пространство знаний без потери детерминированности нижнего контура.

Обмен информацией между уровнями организован через формализованный протокол сообщений, который можно рассматривать как надстройку над инженерной SCADA/DCS-инфраструктурой [11]. LLM-агент получает на вход три категории данных: исторические временные ряды процесса, текущие измерения состояния

и внешние контекстные сведения. Исторические данные используются для построения внутренних векторных представлений состояния системы, аналогичных памяти контекста, обеспечивая осведомлённость агента о предыдущих состояниях регулятора, характерных режимах работы и реакциях на типовые возмущения. Текущие измерения передаются с меньшей частотой, синхронизированной с обновлением RTO, и отражают состояние основных переменных — температуры, мощности, давления и т.д. Контекстная информация поступает через слой интеграции, реализованный на базе Model Context Protocol (MCP), который обеспечивает стандартизированный доступ к файлам, базам данных, API и предсказательным сервисам [10]. MCP не является элементом управления, а выполняет роль интерфейса данных. Он формирует унифицированные запросы к источникам расписаний, погодных моделей, энерготарифов и эксплуатационных журналов, возвращая результаты в формате, пригодном для обработки языковой моделью.

Вся совокупность входных данных агрегируется в текстово-структурированный промпт, который представляет собой машинно читаемое описание текущего состояния и контекста [12]. Типичный промпт включает три блока: (1) обобщённое состояние системы в нормализованной форме с последними измерениями и предсказаниями на ближайший горизонт; (2) набор внешних событий и прогнозов, отражающих операционный контекст; (3) текущие целевые значения, ограничения и коэффициенты весов в критерии MPC, полученные от уровня RTO. В ответ на такой промпт агент формирует текстовый отчёт, содержащий анализ выявленных расхождений между контекстом и текущими целями, а также числовые рекомендации в структурированном формате (например, JSON или XML), где определяются новые значения уставки  $r'$ , весов  $\lambda_y'$ ,  $\lambda_u'$  и допустимых диапазонов  $[y_{\min}', y_{\max}']$ .

Результат работы LLM-агента поступает в модуль RTO, который выполняет машинно читаемую валидацию полученной конфигурации. Проверка осуществляется по критериям достижимости, выполнимости ограничений и совместимости с экономическим функционалом. Если рекомендации удовлетворяют условиям устойчивости и не нарушают физических ограничений процесса, они принимаются и используются в следующей итерации оптимизации. В противном случае инициируется процедура отклонения и возврата к предыдущему профилю. Таким образом, взаимодействие между LLM-агентом и RTO организовано в виде двунаправленного канала, где агент вырабатывает гипотезу реконфигурации, а RTO выполняет её верификацию и формирует обратную связь о допустимости изменений. Эта связь имеет семантический характер, так как отрицательный ответ RTO интерпретируется не как ошибка, а как уточнение границ допустимого множества, что позволяет агенту корректировать последующие рассуждения и повышать точность адаптации.

LLM-агент не функционирует непрерывно в реальном времени, а действует с редким обновлением — обычно на интервале от нескольких минут до часа, соответствуя

масштабу RTO. Такая частота достаточна, поскольку адаптация весов и ограничений не требует высокой дискретизации [2]. Для сохранения вычислительной устойчивости агент работает в асинхронном режиме, не блокируя решение задач оптимизации. Обмен с RTO осуществляется через промежуточный слой очередей сообщений, а обновления MPC производятся только после подтверждения новой конфигурации.

Использование MCP-сервера обеспечивает доступ к множеству гетерогенных источников данных без необходимости ручного кодирования интеграций. Каждый источник описывается в виде контекстного дескриптора, а LLM получает возможность формулировать запросы естественным языком в структурированный формат [10]. Это позволяет использовать как исторические архивы, так и внешние прогнозные сервисы в едином семантическом пространстве. Таким образом, агент способен учитывать долговременные тренды через исторические ряды, краткосрочные события через контекст MCP и текущие измерения через поток данных от RTO/MPC.

С точки зрения функциональной роли LLM-агент представляет собой семантический надзорный уровень, который вводит в контур управления способность интерпретировать внешние знания и контекст. Его выход не является управляющим воздействием в физическом смысле, но задаёт интеллектуальную конфигурацию регулятора. Такой подход обеспечивает согласование стратегического и тактического уровней без необходимости изменять структуру существующих алгоритмов RTO и MPC. Надзорный агент тем самым реализует когнитивную надстройку над инженерным контуром, сохраняя все свойства устойчивости и выполнимости, характерные для классической иерархии управления, но расширяя её возможностью адаптироваться к неформализованным внешним условиям.

## V. ИНТЕГРАЦИЯ LLM-АГЕНТА В ИЕРАРХИЮ RTO–MPC

Интеграция надзорного агента в существующую иерархию управления выполняется с сохранением детерминированной структуры уровней. LLM-агент не вмешивается в внутренние оптимизационные процедуры RTO или MPC, а функционирует в режиме асинхронного контурного взаимодействия. Его роль заключается в периодическом обновлении параметров оптимизационных задач на основании анализа контекста и исторических закономерностей. Временной шаг агента обычно синхронизирован с интервалом обновления RTO, что позволяет интерпретировать внешние данные с горизонтами прогноза от нескольких минут до нескольких часов, не нарушая циклы предиктивного управления.

Для обеспечения формализованного обмена данными между уровнями применяется текстово-структурированный интерфейс, в котором LLM получает промпты в формате JSON [12]. Такой формат позволяет совмещать численные данные, текстовые описания и метаинформацию, обеспечивая воспроизводимость и прозрачность рассуждений. Пример промпта,

передаваемого агенту, приведён ниже:

```
{ "system_state": {
  "timestamp": "2025-10-31T18:00:00Z",
  "temperature_surface": -4.7,
  "temperature_target": -5.0,
  "power_compressor": 0.72,
  "control_signal_u": 0.68,
  "measured_disturbance": 0.05,
  "event_flag": "ice_maintenance"
},
"rto_parameters": {
  "current_setpoint": -5.0,
  "y_min": -5.5,
  "y_max": -4.5,
  "lambda_y": 1.0,
  "lambda_u": 0.3
},
"context": {
  "schedule": "Hockey match starts in 45 minutes",
  "weather_forecast": "outdoor temperature rising from -3°C to +2°C",
  "tariff_level": "high",
  "operator_note": "increase cooling rate before audience arrival"
},
"objective": "analyze current operational context and propose adjusted MPC parameters maintaining energy efficiency and ice quality"
}
```

Агент интерпретирует этот промпт как многомодальное описание состояния: численные данные представляют процесс, а текстовые блоки — внешние события и прогнозы. Результатом работы агента является структурированный ответ, включающий рекомендации для RTO и MPC:

```
{
"recommendations": {
  "new_setpoint": -5.3,
  "lambda_y": 1.2,
  "lambda_u": 0.25,
  "y_min": -5.6,
  "y_max": -4.8,
  "justification": "upcoming match and increasing ambient temperature require deeper cooling for surface stability"
},
"confidence_score": 0.91,
"feasibility_check": "pending"
}
```

Полученный ответ поступает в RTO, где выполняется автоматическая проверка достижимости. Если изменённые параметры не нарушают физические и экономические ограничения, они принимаются и передаются в MPC для обновления конфигурации регулятора.

В формализованном виде алгоритм функционирования надзорного агента можно описать следующим образом:

## Алгоритм 1. Надзорный LLM-Agent интегрированный в RTO-MPC

Input:

- System measurements  $x(t)$ ,  $y(t)$
- RTO configuration parameters
- External context data via MCP (files, APIs, forecasts)
- Historical time-series archive

- 1: Collect current state  $\{x(t), y(t), u(t-1)\}$
- 2: Retrieve contextual information through MCP interface
- 3: Construct structured prompt  $P = \{\text{system\_state}, \text{rto\_parameters}, \text{context}\}$
- 4: Submit  $P$  to LLM-agent for reasoning
- 5: Receive structured recommendations  $R = \{r', \lambda_{y'}, \lambda_{u'}, [y_{\min}', y_{\max}']\}$
- 6: Validate  $R$  on RTO model:
  - if FeasibilityCheck( $R$ ) = True:
    - Apply  $R$  to RTO and forward to MPC
  - else:
    - Reject and revert to previous configuration
- 7: Store ( $P$ ,  $R$ , feedback) in historical log for continuous fine-tuning
- 8: Repeat after  $\Delta T$  (synchronized with RTO update period)

Такой подход обеспечивает модульную интеграцию без изменения существующих интерфейсов управления. Агент использует MCP-серверы как слой контекстной интеграции, позволяющий получать структурированные данные из внешних источников — расписаний, прогнозов, тарифов, журналов обслуживания и исторических архивов — через унифицированные запросы. Это исключает необходимость ручного кодирования API-интеграций и делает LLM-агента восприимчивым к изменению информационной среды без модификации архитектуры [10].

Использование исторических данных в промптах играет роль аналогичной обучению памяти: агент не выполняет статистическое обучение в реальном времени, но опирается на агрегированные характеристики прошедших состояний — например, типичные профили температур и мощности в аналогичных условиях [13]. Таким образом, LLM становится элементом когнитивной реконфигурации, обеспечивая связь между инженерными моделями управления и семантическим контекстом эксплуатации [14].

В предлагаемой архитектуре все решения агента рассматриваются как гипотезы, которые проходят машинно читаемую верификацию на уровне RTO. Этот механизм обеспечивает безопасность, предсказуемость и совместимость гибридной системы «RTO–LLM–MPC» с требованиями промышленной надёжности.

## VI. ОГРАНИЧЕНИЯ, РИСКИ И ПОТЕНЦИАЛ ПРИМЕНЕНИЯ LLM-АГЕНТА

Включение языковых моделей в иерархию управления открывает новые возможности интерпретации контекста и адаптации целей, но одновременно влечёт за собой ряд ограничений, связанных с природой языковых моделей, особенностями промышленной интеграции и требованиями к надёжности систем управления.

Ключевая особенность надзорного агента заключается в том, что он не является автономным звеном, а функционирует как когнитивный слой в окружении строго детерминированных алгоритмов. Поэтому центральным требованием при его применении является сохранение свойств устойчивости и выполнимости RTO–MPC контура при любых сценариях реконфигурации.

Основной источник риска связан с вероятностью ошибочной интерпретации контекста, когда LLM формирует рекомендации, не полностью соответствующие физическим или экономическим ограничениям процесса. Такие ошибки могут возникать из-за неоднозначных текстовых данных, некорректных прогнозов или отсутствия единообразия в источниках MCP-интерфейса [15]. Для минимизации этих рисков вводится процедура машинно читаемой валидации, при которой каждое решение агента проверяется на достижимость и устойчивость средствами RTO. Таким образом, LLM не обладает прямым доступом к исполнительным механизмам — его предложения рассматриваются как гипотезы, проходящие обязательное подтверждение в формализованной модели.

Вторым фактором неопределённости выступают задержки и асинхронность обмена между слоями. Поскольку агент работает с редким обновлением и обрабатывает большие объёмы данных, возникает вероятность временного рассогласования между контекстом и актуальным состоянием процесса. Для нейтрализации этого эффекта взаимодействие организуется через очередь сообщений и буфер временных отметок, обеспечивающий согласование всех данных по времени. При этом регулятор MPC продолжает функционировать автономно на основе последней подтверждённой конфигурации, что исключает влияние задержек на контур управления в реальном времени.

Серьёзное внимание требует и проблема качества контекстных данных, поступающих через MCP-серверы. Ошибки синхронизации, неполные архивы, некорректные метаданные или пропуски прогнозов могут привести к снижению информативности промптов. Для обеспечения устойчивости работы агента вводится система доверительных коэффициентов источников, позволяющая фильтровать контекст и формировать итоговый промпт с весами надёжности [10]. Исторические данные также подлежат нормализации и агрегированию, чтобы не перегружать входное пространство LLM и сохранить воспроизводимость рекомендаций [13].

Несмотря на перечисленные ограничения, потенциал предложенной архитектуры является значительным. LLM-агент обеспечивает впервые реализуемую возможность семантической адаптации регуляторов, при которой изменения уставок и весов в функционале происходят не по фиксированным правилам, а в зависимости от осмысленного контекста эксплуатации. В энергетических и холодильных системах это открывает путь к контекстно-зависимой оптимизации, способной учитывать такие внешние факторы, как расписания

событий, режимы обслуживания, изменения тарифов и погодные колебания. Такая гибкость дополняет традиционную модель RTO–MPC, внося в неё элементы когнитивного управления, при которых решение формируется не только на основе числовых данных, но и на основе семантически интерпретированной информации.

Особый научный интерес представляет перспектива интеграции LLM-агента с предсказательными моделями временных рядов и физико-ориентированными симуляторами. В таком контуре агент может выполнять роль онлайн-медиатора между прогностическими сервисами и оптимизационным ядром, адаптируя горизонты прогнозирования и критерии в зависимости от контекста. Это позволит реализовать гибридные когнитивно-инженерные системы управления, в которых машинное рассуждение дополняет математическую оптимизацию. Применение подобных подходов в системах охлаждения, кампусных энергетических центрах и интеллектуальных зданиях способно обеспечить экономию энергии, повышение устойчивости регулирования и снижение потребности в ручной перенастройке.

В долгосрочной перспективе LLM-агент может рассматриваться как универсальный интерфейс интерпретации контекста в инженерных системах, обеспечивающий симбиоз языковых технологий и методов оптимального управления. При этом сохранение формальных гарантий устойчивости достигается за счёт чёткого разграничения уровней: языковой агент — это аналитический модуль, а принятие решений остаётся в зоне ответственности формализованных алгоритмов RTO и MPC. Такая архитектура создаёт основу для нового поколения систем управления, сочетающих предсказательную строгость и когнитивную гибкость.

## VII. ЗАКЛЮЧЕНИЕ

Представленная работа формирует концептуальную основу интеграции языковых моделей в иерархию оптимального управления энергетическими и холодильными системами. Предложена архитектура надзорного уровня, в которой агент на базе большой языковой модели (LLM) функционирует между слоями экономической оптимизации (RTO) и модельного предиктивного управления (MPC), обеспечивая когнитивное согласование целей и ограничений с внешним контекстом эксплуатации. Архитектура сохраняет математическую строгость классической схемы RTO–MPC, но расширяет её возможностями семантической адаптации, позволяя учитывать текстовые и неструктурированные источники информации.

Разработанная схема взаимодействия, включающая формат промптов, интерфейс MCP и алгоритм машинно читаемой валидации, обеспечивает безопасную интеграцию агента без вмешательства в контур реального времени. Все решения агента рассматриваются как гипотезы и проходят проверку на достижимость и устойчивость средствами RTO, что гарантирует сохранение инженерных свойств системы.

Таким образом, LLM становится когнитивным надзорным слоем, способным интерпретировать контекст, прогнозировать изменения эксплуатационных условий и реконфигурировать параметры регулятора при сохранении формальных гарантий выполнимости.

Потенциал предложенного подхода выходит за пределы конкретного применения в холодильных системах. Включение когнитивных агентов в контур управления открывает путь к созданию гибридных архитектур, объединяющих строгие методы оптимизации и семантические механизмы рассуждения. Такие системы способны к адаптации в условиях неопределённости, снижению потребления энергии, улучшению предсказуемости поведения и сокращению необходимости ручного вмешательства. Дальнейшие исследования могут быть направлены на формализацию интерфейсов взаимодействия между LLM и физическими моделями, а также на изучение устойчивости когнитивно-инженерных контуров в промышленных условиях.

## БИБЛИОГРАФИЯ

- [1] Vaccari M., Pannocchia G. A modifier-adaptation strategy towards offset-free economic MPC // *Processes*. – 2016. – Т. 5. – №. 1. – С. 2.
- [2] Afram A., Janabi-Sharifi F. Theory and applications of HVAC control systems—A review of model predictive control (MPC) // *Building and environment*. – 2014. – Т. 72. – С. 343-355.
- [3] Serale G. et al. Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities // *Energies*. – 2018. – Т. 11. – №. 3. – С. 631.
- [4] Darby M. L. et al. RTO: An overview and assessment of current practice // *Journal of Process control*. – 2011. – Т. 21. – №. 6. – С. 874-884.
- [5] Hovgaard T. G. et al. Model predictive control technologies for efficient and flexible power consumption in refrigeration systems // *Energy*. – 2012. – Т. 44. – №. 1. – С. 105-116.
- [6] Bujarbaruah M. et al. Adaptive MPC for iterative tasks // *2018 IEEE Conference on Decision and Control (CDC)*. – IEEE, 2018. – С. 6322-6327.
- [7] Wu R., Ai J., Li T. InstructMPC: A Human-LLM-in-the-Loop Framework for Context-Aware Control // *arXiv preprint arXiv:2504.05946*. – 2025.
- [8] Tøndel P., Johansen T. A., Bemporad A. An algorithm for multi-parametric quadratic programming and explicit MPC solutions // *Automatica*. – 2003. – Т. 39. – №. 3. – С. 489-497.
- [9] Huang X. et al. Understanding the planning of LLM agents: A survey // *arXiv preprint arXiv:2402.02716*. – 2024.
- [10] Ahmadi A., Sharif S., Banad Y. M. MCP Bridge: A Lightweight, LLM-Agnostic RESTful Proxy for Model Context Protocol Servers // *arXiv preprint arXiv:2504.08999*. – 2025.
- [11] Adamo F. et al. SCADA/HMI systems in advanced educational courses // *IEEE Transactions on Instrumentation and measurement*. – 2007. – Т. 56. – №. 1. – С. 4-10.
- [12] White J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt // *arXiv preprint arXiv:2302.11382*. – 2023.
- [13] Jin M. et al. Time-llm: Time series forecasting by reprogramming large language models // *arXiv preprint arXiv:2310.01728*. – 2023.
- [14] Wang X. et al. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection // *Advances in Neural Information Processing Systems*. – 2024. – Т. 37. – С. 58118-58153.
- [15] Ji Z. et al. Towards mitigating LLM hallucination via self reflection // *The 2023 Conference on Empirical Methods in Natural Language Processing*. – 2023.

# Intelligent Supervisory Layer Based on a Language Model for Real-Time Optimization Systems

**Abstract**— This paper presents a conceptual framework for integrating large language models (LLMs) into hierarchical industrial control architectures based on Real-Time Optimization (RTO) and Model Predictive Control (MPC). The proposed supervisory LLM layer serves as a cognitive interface between economic optimization and predictive regulation, providing context-aware adaptation of setpoints, constraints, and cost function weights. A structured communication protocol is introduced, enabling the agent to receive system measurements, historical time-series data, and external contextual information through a Model Context Protocol (MCP) interface. The LLM interprets these heterogeneous inputs and generates machine-readable recommendations for the RTO layer, which validates them through feasibility and stability checks before application. This ensures that cognitive reasoning remains safely encapsulated within the formal guarantees of hierarchical control. The resulting architecture maintains the mathematical rigor of RTO-MPC while extending its adaptability to unstructured knowledge sources, such as schedules, forecasts, and operational reports. The approach enables semantic adaptation of control parameters without compromising real-time performance and offers potential for implementation in energy management, refrigeration systems, and intelligent building operations.

**Keywords** — RTO-MPC, LLM agent, hierarchical control, cognitive optimization, semantic adaptation, MCP interface, predictive regulation.

## REFERENCES

- [1] Vaccari M., Pannocchia G. A modifier-adaptation strategy towards offset-free economic MPC //Processes. – 2016. – T. 5. – №. 1. – C. 2.
- [2] Afram A., Janabi-Sharifi F. Theory and applications of HVAC control systems—A review of model predictive control (MPC) //Building and environment. – 2014. – T. 72. – C. 343-355.
- [3] Serale G. et al. Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: Problem formulation, applications and opportunities //Energies. – 2018. – T. 11. – №. 3. – C. 631.
- [4] Darby M. L. et al. RTO: An overview and assessment of current practice //Journal of Process control. – 2011. – T. 21. – №. 6. – C. 874-884.
- [5] Hovgaard T. G. et al. Model predictive control technologies for efficient and flexible power consumption in refrigeration systems //Energy. – 2012. – T. 44. – №. 1. – C. 105-116.
- [6] Bujarbaruah M. et al. Adaptive MPC for iterative tasks //2018 IEEE Conference on Decision and Control (CDC). – IEEE, 2018. – C. 6322-6327.
- [7] Wu R., Ai J., Li T. InstructMPC: A Human-LLM-in-the-Loop Framework for Context-Aware Control //arXiv preprint arXiv:2504.05946. – 2025.
- [8] Tøndel P., Johansen T. A., Bemporad A. An algorithm for multi-parametric quadratic programming and explicit MPC solutions //Automatica. – 2003. – T. 39. – №. 3. – C. 489-497.
- [9] Huang X. et al. Understanding the planning of LLM agents: A survey //arXiv preprint arXiv:2402.02716. – 2024.
- [10] Ahmadi A., Sharif S., Banad Y. M. MCP Bridge: A Lightweight, LLM-Agnostic RESTful Proxy for Model Context Protocol Servers //arXiv preprint arXiv:2504.08999. – 2025.
- [11] Adamo F. et al. SCADA/HMI systems in advanced educational courses //IEEE Transactions on Instrumentation and measurement. – 2007. – T. 56. – №. 1. – C. 4-10.
- [12] White J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt //arXiv preprint arXiv:2302.11382. – 2023.
- [13] Jin M. et al. Time-llm: Time series forecasting by reprogramming large language models //arXiv preprint arXiv:2310.01728. – 2023.
- [14] Wang X. et al. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection //Advances in Neural Information Processing Systems. – 2024. – T. 37. – C. 58118-58153.
- [15] Ji Z. et al. Towards mitigating LLM hallucination via self reflection //The 2023 Conference on Empirical Methods in Natural Language Processing. – 2023.