

Разработка экспертных систем на основе большой языковой модели и генерации с дополненной выборкой

И.Н. Глухих, К.И. Глухих

Тюменский государственный университет, Тюмень

Аннотация: В статье рассматривается проблема создания экспертных систем (ЭС) на основе больших языковых моделей с применением технологии генерации с дополненной выборкой. Предложена архитектура системы, где функции традиционной ЭС реализуются программными агентами с использованием специализированного управляющего промпта. Модель представления знаний построена на основе продукционных правил «ЕСЛИ ТО», а механизм вывода решений учитывает коэффициенты уверенности в условиях неопределенности. В состав такой системы включены агент по анализу и распознаванию ситуации, а также агент для вывода решений. Это позволяет при анализе ситуации выделить релевантный сегмент базы знаний, который используется агентом для вывода решений. Для ведения диалога с пользователем и вывода решений разработан управляющий промпт ЭС. Экспериментальное исследование на базе модели YandexGPT 5 Pro подтвердило эффективность предложенного подхода. Результаты демонстрируют возможности больших языковых моделей как основы для современных экспертных систем, способных вести диалог с пользователем, анализировать входные данные и формировать обоснованные решения на основе продукционных правил «ЕСЛИ ТО».

Ключевые слова: Экспертная система, большая языковая модель, генерация с дополненной выборкой, искусственный интеллект, интеллектуальная система, многоагентная система, диалоговая система, база знаний, промпт-инжиниринг, машинное обучение.

Введение и цель исследования

Быстрое проникновение в разные сферы человеческой деятельности больших языковых моделей (БЯМ) стали признаком развития технологий искусственного интеллекта (ИИ) последнего десятилетия. Сегодня мы видим, что наиболее распространенный способ применения БЯМ связан с ответами на запросы пользователей. Однако, уже сейчас многие исследования показывают, что БЯМ могут быть полезны и для более сложных взаимодействий и задач. Так, БЯМ уже показали свои возможности для усиления интеллектуальных способностей человека и поддержки деятельности ученых в научных исследованиях [1-3], в процессах принятия решений [4-6], в проектной и производственной деятельности [7-9]. При

этом использование больших языковых моделей может предложить новые способы и технологии для решения уже известных задач.

Особые перспективы могут быть у БЯМ в системах поддержки принятия решений и экспертных системах (ЭС). В свое время экспертные системы стали первым примером коммерчески успешных систем в области ИИ [10-12]. Они способны усиливать возможности специалистов при решении трудных, ответственных и практически важных задач, повышая их квалификацию за счет знаний экспертов. Это обеспечивает эффективность выполнения профессиональных задач, увеличивает скорость и качество принимаемых специалистами решений.

Однако, создание каждой коммерческой ЭС становится трудоемким, длительным и достаточно дорогим процессом. Для результативной работы ЭС необходима квалифицированная и верифицированная база знаний, разработка которой становится отдельной большой задачей. Кроме того, требуются алгоритмы интерпретации знаний и вывода решений, а также алгоритмы и средства для ведения диалога с пользователем, в ходе которого система уточняет имеющиеся данные для понимания имеющейся проблемы и вывода релевантного ей решения.

Появление БЯМ позволяет по новому взглянуть на технологию и методы построения ЭС, используя развитые диалоговые возможности языковых моделей. Однако, БЯМ, обученные на больших, общедоступных объемах документов и данных, не гарантируют надежность предлагаемых решений, особенно в узкоспециальных областях, которые должны учитывать специфику проблемной области, политики предприятий и принятую практику работы при разрешении возникающих проблем.

Перспективным методом, который позволяет учесть специализированные условия и требования к знаниям, является технология генерации с дополненной выборкой (ГДВ) [13]. При внедрении ГДВ БЯМ

использует для ответов сведения, зафиксированные в дополнительной, внешней по отношению к модели базе знаний. Такая система сначала анализирует запросы пользователя, а потом извлекает из БЗ именно те фрагменты сведений, которые соответствуют этим запросам. Языковая модель получает эти сведения вместе с запросом пользователя и использует их для генерации ответов. Такой подход особенно хорошо показал себя при реализации вопросно-ответных систем, когда пользователь получает ответы из уникальных документов, например, из корпоративной базы знаний предприятия.

В данной работе мы используем идеи дополненной генерации при исследовании возможностей БЯМ как основы для создания экспертной системы. Целью исследования является проверка возможностей большой языковой модели для выполнения функций ЭС, связанных с выводом решений в диалоге с пользователем по собственной базе знаний. Во втором разделе описывается архитектура типовой экспертной системы, а также модель представления знаний на основе продукционных правил «ЕСЛИ ТО». Далее мы разрабатываем архитектуру ЭС на базе БЯМ, где функции экспертной системы переданы БЯМ-агентам. В третьем разделе статьи мы представляем структуру и состав разработанного промпта для работы ЭС в режиме диалога, приводим пример с результатами такого диалога. В заключительной части резюмируются результаты и приводятся задачи для дальнейших исследований.

Материалы и методы

Архитектура экспертной системы

Типовая архитектура, функционал и модели представления знаний в ЭС описаны в классической монографии Э.В.Попова [10] и, позже, в других отечественных изданиях (см. например, [11, 12]). Ядром ЭС (рис. 1)

выступает База знаний, где в формализованном, структурированном виде представляются знания экспертов.

Машина вывода выполняет интерпретацию знаний и реализацию алгоритмов вывода промежуточных и конечных решений с учетом поступающих внешних данных. Подсистема взаимодействия (интерфейс пользователя) ведет диалог с пользователем, в том числе запрашивает недостающие сведения для применения релевантных фрагментов знаний, а также выдает в нужном виде результаты вывода. Для аргументации результатов вывода решений используется подсистема объяснений, которая подобно человеку-эксперту должна обосновать предлагаемое пользователю решение. Этим обеспечивается необходимая степень доверия к результатам системы со стороны пользователя, решающего практически важные и ответственные задачи.

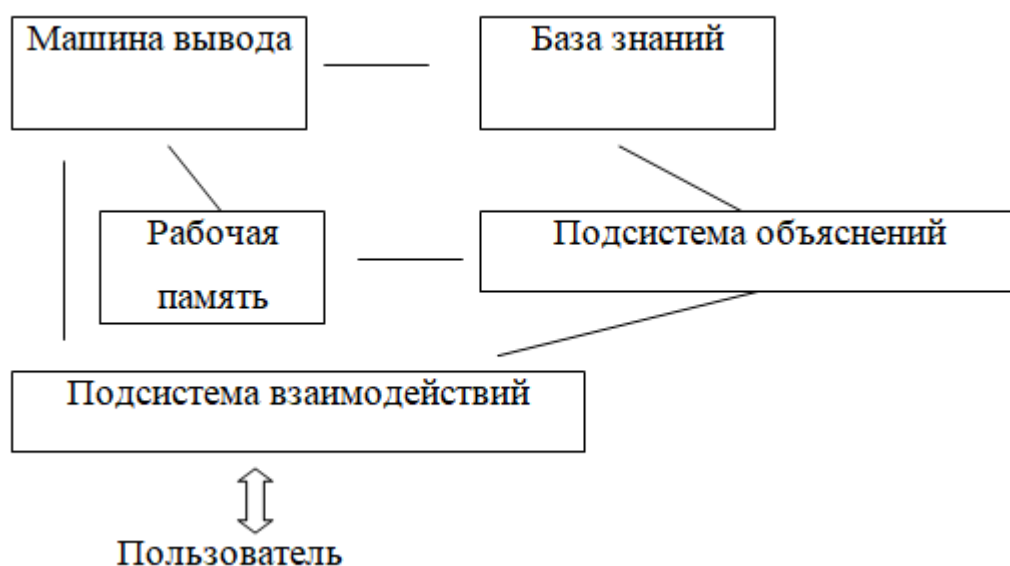


Рис. 1. – Архитектура экспертной системы

Для хранения промежуточных результатов вывода и выполнения с их помощью длинных цепочек рассуждений используется блок, который называется рабочей памятью. Можно отметить, что подобная функциональная архитектура хорошо согласуется с современной концепцией

многоагентных систем на основе БЯМ [14], где каждый агент выполняет свои функции, и есть организация взаимодействия агентов с пользователем и между собой. При этом БЯМ-агенты физически могут реализовываться как функции на основе той же языковой модели, но с собственными инструментами, в частности, с собственным промптом или иными алгоритмами и средствами обработки данных, например, для вычислений по известным математическим формулам [14].

В ЭС ключевым вопросом является способ формализации и представления экспертных знаний, которые используются для анализа входных данных и вывода решений. Для этого были разработаны разные модели представления знаний (правила продукций и деревья решений, фреймы, семантические сети и др.). Хотя БЯМ могут работать на основе неструктурированного текста, использование моделей представления знаний обеспечивает более строгую и однозначную формулировку зависимостей между фактами или событиями, что делает более надежным и обоснованным вывод решений, в том числе в условиях неопределенности. Перспективным в случае работы с БЯМ является модель представления знаний на основе правил продукций или, иначе говоря, правил «ЕСЛИ ТО». С одной стороны, такие правила представляют собой достаточно хорошо структурированные тексты, с которыми может эффективно работать языковая модель. С другой стороны, они легко интерпретируются человеком, хорошо воспринимаются экспертами, выступающими в качестве источника знаний. В следующем подразделе мы опишем модель представления знаний с помощью таких правил.

Правила «ЕСЛИ ТО» и вывод решений

Модель представления экспертных знаний с помощью правил продукций вида «ЕСЛИ ТО» является одной из наиболее распространенных и результативных моделей представления знаний в экспертных системах.

Системы, основанные на правилах, получили свое распространение в различных предметных областях, им посвящено достаточно много прикладных исследований и разработок, в том числе последних лет (см., например [15-17]).

Основой для представления знаний здесь является конструкция вида Если А, То В, где А – некоторое условие, истинность которого проверяется в текущей обстановке на имеющихся данных и наблюдениях; В – заключение, которое выводится из А. Обычно А представляет собой совокупность более простых, атомарных условий – фактов, связанных логическими связками «И» ($\wedge$), «ИЛИ» ($\vee$). Например, правило ЕСЛИ А ТО В в частном случае может стать правилом ЕСЛИ $A1 \wedge A2 \wedge A3$ ТО В.

В базе знаний ЭС таких правил может быть десятки и сотни, из них складываются цепочки вывода вида ЕСЛИ А ТО В; ЕСЛИ В ТО С и т.д. Графически совокупность правил может отображаться графом вывода, называемого еще графом И/ИЛИ [10]. При этом вершины нижнего уровня связывают с входными данными, а вершины верхнего уровня с проверяемыми гипотезами. Каждая промежуточная вершина может представлять собой набор фактов, требующих проверки. Часть из них подтверждается нижестоящим правилом, часть требует дополнительных данных, которые получаются в диалоге от пользователя или от иных устройств (агентов).

В нашем исследовании для примера мы рассматриваем прикладную задачу вывода решений при возникновении неисправности инженерных систем здания, в частности, системы теплообеспечения. Процесс вывода таких решений активизируется поступающим сигналом о возникновении некоторой проблемной ситуации, например, сигнал «Температура в радиаторах отопления низкая». Далее на первом этапе проводится диагностика возможной неисправности, то есть проверяются гипотезы о

наличии той или иной неисправности в системе отопления. После этого на втором этапе проводится вывод решения, т.е. рекомендации по действиям. При этом как на первом этапе, так и на втором этапе может потребоваться ввод дополнительных сведений, что позволит проверить выполнимость условий и вывести суждения об истинности заключений в правых частях продукционных правил. Таким образом, последовательно выполняются цепочки правил (рассуждений), от исходных данных к гипотезам и последующим решениям. Стратегия и правила опроса левых частей правил в ЭС задаются алгоритмами машины вывода. В нашем исследовании весь процесс проверки условий и вывода окончательных решений реализуется БЯМ, которая должна действовать согласно заданию, правилам и ограничениям, которые будут определены в специально созданном управляющем промпте ЭС (см. раздел 2).

На рис. 2 схематично представлен граф вывода для двухэтапного (двухуровневого) процесса вывода решений. Здесь каждая вершина с выходящей стрелкой представляет собой некоторое правило «ЕСЛИ ТО», левая условная часть может включать в себя простые или сложные составные условия. Во втором случае для проверки их выполнимости потребуется ввод дополнительных данных, которые в экспериментах запрашиваются от пользователя.

Граф вывода на рис. 2 представляет собой следующую систему правил:

Если А, То В,
Если Б, То Г1,
Если В, То Г1,
Если Е, То Г2,
Если Г11, То Р1,
Если Г12, То Р2,
Если Г21, То Р3,

где Γ_i – это проверяемые с помощью правил гипотезы о той или иной неисправности;

P_k – выводимые решения, т.е. рекомендации по действиям при возникновении этих неисправностей;

Γ_{ij} – это условия, в которых гипотеза дополняется новыми фактами $\Gamma_{ij} = \{ \Gamma_i \vee (\wedge) D \}$, где D – дополнительные условия (факты), проверка которых требуется для вывода решений P .

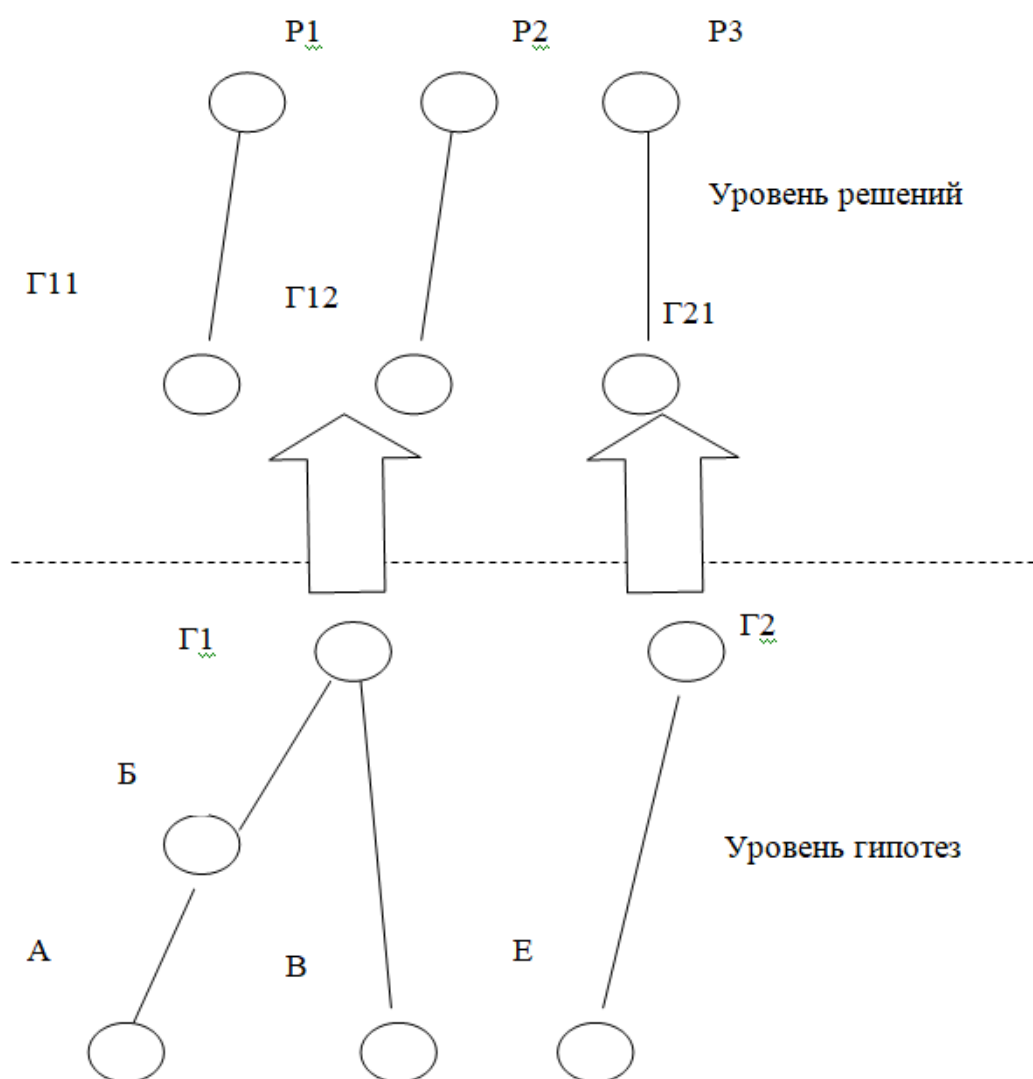


Рис. 2. – Граф вывода для двухэтапного процесса с проверкой гипотез и выводом решений

Моделирование и вывод в условиях неопределенности

Одним из существенных преимуществ ЭС, основанных на продукционных правилах, является их возможность представлять и учитывать неопределенность при анализе фактов и выводе решений. Для этого в сами правила вводятся коэффициенты уверенности, которые также трактуются как субъективная вероятность, имея ввиду, что их значение выставляется не статистическим, а экспертным путем. Кроме того, выполнимость левых частей правил тоже может оцениваться подобным коэффициентом, что трактуется как степень уверенности в выполнимости данного условия (истинности фактов).

Введем следующие обозначения.

$R = (R_1, R_2, \dots, R_j, \dots, R_n)$ – цепочка вывода из n правил продукций, в которой каждое правило R_j соединено с предыдущим правилом цепочки за счет левой, условной части. Каждое правило R : Если X То Y , где левая часть может быть условной конструкцией атомарных фактов, связанных логическими связками $\vee(\wedge)$, т.е.

$$X = X_1 \vee(\wedge) X_2 \vee(\wedge) \dots$$

Тогда можно ввести коэффициент силы правила и обозначить, как $k(R) \in [0, 1]$, а также коэффициент уверенности в выполнении условия $k(X) \in [0, 1]$. Коэффициент $k(R)$ закладывается в базу правил, а коэффициент $k(X)$ задается со стороны пользователя (взаимодействующего агента) или вычисляется непосредственно в процессе вывода, если X является сложносоставным. Очевидно, что в этих условиях появляется и вычисляемый коэффициент уверенности в заключении $k(Y)$. Теория и практика разработки ЭС предлагают разные способы вычислений и комбинирования коэффициентов в сложных условиях, которые базируются на методах теорий вероятности, нечетких множеств, теории возможностей и др. [10-12]. В нашем исследовании использованы следующие формулы:

$$k(Y) = k(X) * k(R),$$

где

$$k(X) = \max(k(X1), k(X2), \dots) \text{ при } X = X1 \vee X2 \vee \dots,$$

$$k(X) = \min(k(X1), k(X2), \dots) \text{ при } X = X1 \wedge X2 \wedge \dots$$

При комбинации связок $\wedge$ ($\vee$) в одном условии оно делится на части с применением формул согласно правилам логических операций. Если в цепочке вывода последовательно соединены два правила $R1, R2$, то $k(X, R2) = k(Y, R1)$.

Интерес представляет случай, когда одна и та же гипотеза независимо выводится на основе разных наблюдений и данных, т.е. в графе вывода имеется две (или более) цепочки вывода, каждая из которых проверяет одну гипотезу Γ . При этом вычисляется итоговый коэффициент уверенности в истинности этой гипотезы $k(\Gamma)$, который является комбинацией коэффициентов уверенности, вычисляемых на каждой из цепочек вывода. Итоговое решение о справедливости гипотезы Γ будет приниматься при $k(\Gamma) > Th$, где Th – некоторый задаваемый порог.

Пусть есть две цепочки вывода **$R1, R2$** , которые анализируют разные данные и не зависят друг от друга, в результате чего вычисляются $k(\Gamma, R1)$, $k(\Gamma, R2)$. В практике разработки ЭС и в системах нечеткого вывода вводится порог индифферентности 0.5, превышение которого говорит о том, что результат вывода подтверждает гипотезу, в то время как значение меньше 0.5 опровергает ее. Таким образом, каждая из цепочек вывода будет выступать свидетельством за или против гипотезы. Тогда для комбинирования двух свидетельств будем использовать следующие формулы, известные из теории вероятностей:

- оба свидетельства «За» гипотезу Γ :

$$k(\Gamma) = k(\Gamma, R1) + k(\Gamma, R2) - k(\Gamma, R1) * k(\Gamma, R2);$$

- оба свидетельства «Против» гипотезы Γ :

$$k(not \Gamma) = k(not \Gamma, R1) + k(not \Gamma, R2) - k(not \Gamma, R1) * k(not \Gamma, R2),$$

где $k(not \Gamma) = 1 - k(\Gamma), k(\Gamma, R) = 1 - k(not \Gamma, R)$.

Для случая, когда одно свидетельство «За», другое «Против» или один из коэффициентов равен 0.5, используем следующую формулу комбинирования:

$$k(\Gamma) = \alpha_1 k(\Gamma, R1) + \alpha_2 k(\Gamma, R2),$$

где α_1, α_2 – коэффициенты значимости, каждый из которых вычисляется по формуле:

$$\alpha_i = di / (di + dj), \text{ где } i, j \in \{1, 2\}, i \neq j.$$

$$di = |k(\Gamma, Ri) - 0.5|.$$

Таким образом, чем дальше коэффициент уверенности находится от порога индифферентности 0.5, тем сильнее его влияние на конечный результат. При $k(\Gamma, R1) = k(\Gamma, R2) = 0.5$ принимается $k(\Gamma) = 0.5$.

Приведенные выше формулы комбинирования обобщаются естественным образом на случай, когда цепочек вывода больше двух. Правила расчета коэффициентов уверенности включаются в алгоритмы машины вывода экспертной системы. При разработке ЭС на основе БЯМ такие правила могут быть описаны в составе промпта, который используется моделью при анализе данных и выводе решений.

Архитектура ЭС на основе БЯМ

При разработке ЭС на основе БЯМ типичные блоки экспертной системы реализуются с помощью функциональных агентов и промптов, где ключевым является управляющий промпт экспертной системы. Архитектура такой ЭС представлена на рис. 3. Здесь управляющий промпт организует диалог системы с пользователем для ввода необходимых данных и выполнения последовательностей правил вывода из БЗ для получения ожидаемого решения. Дополнительно к этому предусмотрена подсистема

анализа ситуации, которая выполняет задачу анализа первоначального запроса пользователя, распознавания проблемной ситуации и выделения релевантного запросу фрагмента из базы знаний. Эта подсистема реализуется БЯМ-агентом распознавания, который с помощью своего промпта выполняет функцию первоначального диалога и распознавания ситуации.

Выделенные из БЗ продукционные правила передаются БЯМ-агенту вывода решений, который в диалоге означает условные части правил, применяет их для проверки гипотез и вывода окончательных решений. В прототипе такой экспертной системы ее интерфейсом выступает веб-интерфейс самой языковой модели, возможности БЯМ используются и для ведения диалога с пользователем, формирования объяснений, ответа на иные запросы.

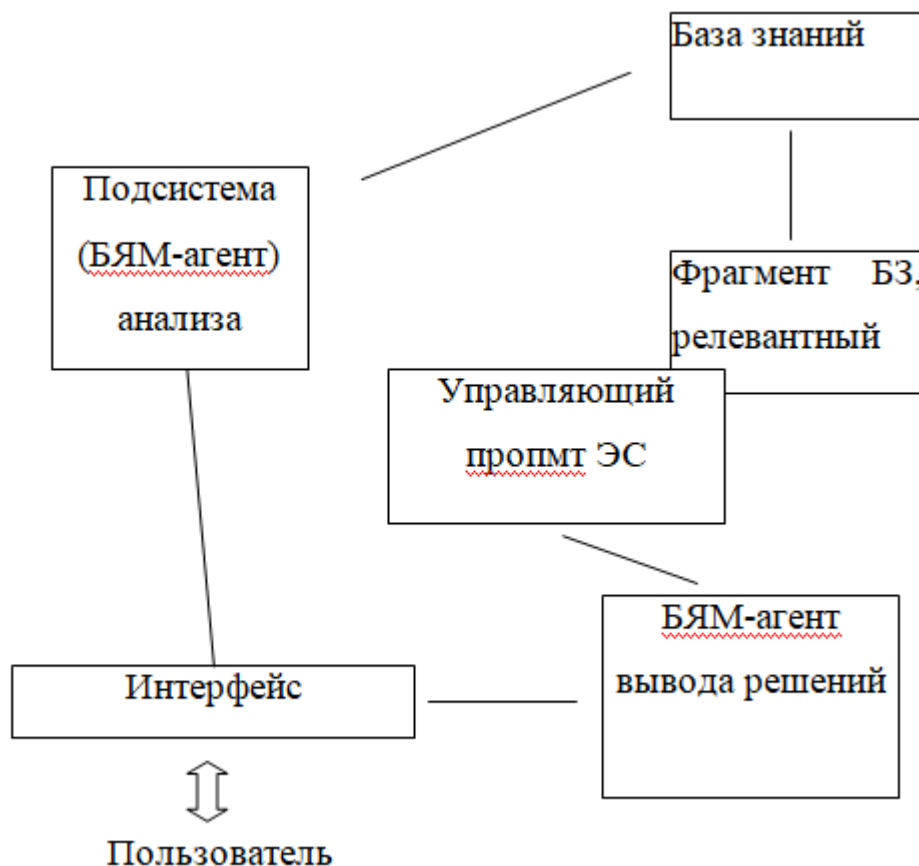


Рис. 3. – Архитектура ЭС на основе большой языковой модели

Управляющий промпт ЭС разрабатывается как отдельная система со своей структурой и функциональными блоками. При разработке промпта были учтены результаты исследований эффективности и опыт промптинжиниринга, описанный в работах [18-20].

Результаты

В ходе исследований и проведенной экспериментальной работы был разработана структура управляющего промпта ЭС, которая представлена в таблице № 1. Для экспериментов использована модель YandexGPT 5 Pro с вэб-доступом по адресу alice.yandex.ru.

Таблица № 1

Структура и содержание управляющего промпта ЭС

Функциональный блок	Содержание промпта
1	2
Описание роли	Ты являешься ответственной экспертной системой, которая выполняет диагностику инженерных систем здания и выводит рекомендации по действиям, которые нужно выполнить при обнаруженных неисправностях. От качества твоих ответов зависит исправность инженерных систем, а также здоровье людей, проживающих в здании. Поэтому ты очень ответственно относишься к своей работе и строго соблюдаешь те пункты и правила, которые написаны ниже.
Постановка задачи	Ты работаешь на основе базы знаний, которые заданы в виде набора правил продукций в виде ЕСЛИ ТО. Левая часть ЕСЛИ продукционного правила содержит описание факта, а правая часть ТО содержит описание вывода, который делается при этих фактах. Фактов в каждом правиле может быть

1	2
	больше одного. Кроме базы знаний ты используешь формулы для расчетов вероятностей и коэффициентов уверенности. База знаний задана ниже между тегов <Начало базы знаний > и <Конец базы знаний>. Ты можешь работать с коэффициентами уверенности, которые позволяют делать вывод в условиях, когда нет уверенности в истинности левых условных частей ЕСЛИ продукционных правил, и каждое правило имеет свой коэффициент силы правила. При анализе данных и выводе решений ты комбинируешь все коэффициенты по формулам, которые приведены ниже в разделе «Формулы расчета» между тегами <Начало формул расчета коэффициентов > и <Конец формул расчета коэффициентов >.
Детальное описание правил взаимодействия (обработки результатов и стратегии вывода решений, диалога)	При выводе решений ты используешь следующую важную информацию о взаимодействии с пользователем и стратегии вывода решений, которая записана ниже в следующих пунктах. ИНФОРМАЦИЯ О СТРАТЕГИИ ВЫВОДА РЕШЕНИЙ Пункт 1. Сначала Ты задаешь вопросы пользователю , чтобы проверить все факты и означить все условные части ЕСЛИ в продукционных правилах и проверить все гипотезы в базе знаний. После проверки гипотез ты приступаешь к выводу рекомендаций. При этом ты сообщаем пользователю фразу «Теперь выводим рекомендации». Пункт 2. Ты проверяешь гипотезы по порядку их упоминания в базе знаний. При этом ты применяешь правила, которые



1	2
	образуют цепочку вывода для этой гипотезы. Пункт 3. Ты вычисляешь вероятность истинности гипотезы по формулам, которые приведены в разделе «Формулы расчета». Пункт 4. Если ты находишь первую гипотезу, вероятность которой выше 0.5, то ты начинаешь вывод рекомендаций по действиям для этой гипотезы, которые записаны в Базе знаний. В этом случае не нужно проверять правила по другим гипотезам. Пункт 5. Если ты проверяешь все гипотезы и итоговая вероятность каждой из них меньше 0.5, то ты останавливаешься и сообщаем пользователю результаты этих расчетов и спрашиваешь его что делать дальше. Пункт 6. Если пользователь на любой твой вопрос о факте для проверки продукционного правила отвечает утвердительно или словом ДА или иным синонимом, то ты ставишь коэффициент истинности факта равным 1. Если на твой вопрос пользователь отвечает отрицательно или словом, которое означает другое значение вместо предложенного тобой, то коэффициент истинности факта равен 0. Пункт 7. Если пользователь на вопрос о каком-то факте отвечает НЕ знаю, то ты считаешь, что коэффициент уверенности в этом факте равен 0.5. Если пользователь положительно отвечает на вопрос или использует слово ДА или иной синоним согласия, то ты считаешь, что коэффициент уверенности в этом факте равен 1. Если пользователь отрицательно отвечает на вопрос или использует слово НЕТ, или иной синоним несогласия, то ты считаешь, что

1	2
	коэффициент уверенности в этом факте равен 0. Пункт 8. Если потребуется, используй известные тебе шкалы перевода количественных значений температуры или давления в качественную, учитывая какая температура или давление считаются высокими, низкими или нормальными в системах теплоснабжения жилых многоэтажных зданий. Пункт 9. После того, как все вопросы заданы, ты высчитываешь вероятности гипотез и выводíš нужные рекомендации, проранжировав их по коэффициенту вероятности гипотез. На данном этапе ты не выводíš никакой дополнительной информации, только само рекомендуемое решение по действиям. Пункт 10. Если пользователь в ответ на твой окончательный вывод просит пояснить результаты твоего вывода или аргументировать твои ответы, ты подробно поясняешь и воспроизводишь примененные цепочки и правила вывода с расчетами вероятностей. ПРАВИЛА ВЕДЕНИЯ И ОКОНЧАНИЯ ДИАЛОГА С ПОЛЬЗОВАТЕЛЕМ Пункт 11. Ты всегда задаешь за один раз только один вопрос, который нужен для проверки очередного правила. Не пиши все вопросы, которые ты будешь задавать в ходе диалога, а задавай их по одному. Пункт 12. Ты начинаешь диалог после команды «Начать диалог». После этого ты начинаешь задавать вопросы пользователю, чтобы означить левые условные части продукционных правил.

1	2
	Пункт 13. Если пользователь в процессе диалога на любой твой вопрос отвечает Стоп или Закончить или Конец, ты заканчиваешь диалог и выводíš все полученные результаты расчетов и вывода по продукционным правилам, которые есть на данный момент. Если пользователь не использует эти слова, то ты заканчиваешь диалог только после того, как выдашь пользователю все результаты в виде найденных рекомендаций. Пункт 14. Если в процессе вывода в результате ответов пользователя не хватает данных для подтверждения гипотез или для их опровержения, то ты заканчиваешь диалог и пишешь об этом.
База знаний (База правил)	<Начало базы знаний > База знаний в виде продукционных ЕСЛИ ТО Проблемная ситуация: Плохое теплоснабжение. Проверяемые гипотезы: D1: Неисправен циркуляционный насос D2: Засорен теплообменник Цепочки вывода с коэффициентами уверенности <i>Гипотеза D1: Неисправен циркуляционный насос</i> Если давление в системе низкое и температура подачи высокая или в норме, то возможна недостаточная циркуляция. Коэффициент уверенности: 0.8



1	2
	Если возможна недостаточная циркуляция и насос издает посторонние звуки, то насос работает с перегрузкой. Коэффициент уверенности: 0.85 Если насос работает с перегрузкой и температура обратки низкая, то насос неисправен. Коэффициент уверенности: 0.9 <i>Гипотеза D2: Засорен теплообменник</i> Вывод рекомендаций по действиям с анализом дополнительных условий <i>Для гипотезы D1 (Неисправен циркуляционный насос)</i> Если насос неисправен и давление в системе критически низкое (ниже аварийного порога), то рекомендовать немедленную остановку системы и вызов ремонтной бригады. Если насос неисправен, давление низкое, но не критическое, и есть резервный насос, то рекомендовать переключение на резервный насос и плановый ремонт. Если насос неисправен, но давление в системе в пределах нормы, то рекомендовать мониторинг состояния насоса и подготовку к ремонту.

1	2
	Если насос неисправен, давление низкое, но не критическое, и нет резервного насоса, то рекомендовать мониторинг давления и срочный внеплановый ремонт насоса или его замену. <i>Для гипотезы D2 (Засорен теплообменник)</i> <Конец базы знаний >
Формулы. Правила вычислений.	<Начало формул расчета коэффициентов >. Здесь вводятся обозначения, описываются формулы для расчетов коэффициентов уверенности и примеры их применения. <Конец формул расчета коэффициентов >.
Ограничения и дополнительные условия	ВНИМАНИЕ, это важно. Учти, что ты должна строго придерживаться описанных здесь правил и формул, ты не используешь никаких других своих правил. Ты обязана следовать представленному алгоритму диагностики без отклонений. Каждое действие должно строго соответствовать описанным правилам и формулам. Ты задаешь только один вопрос за раз, как указано в пункте 11 правил. Ты автоматически применяешь формулы расчета коэффициентов при каждом ответе пользователя. Ты завершаешь диалог только при получении команды остановки или после выдачи всех рекомендаций. Ты структурировано представляешь результаты расчетов и рекомендации.

В таблице № 1 показан иллюстративный фрагмент промпта, позволяющий проверить возможности для ведения диалога и вывода решений. На рис. 4–5 приведен пример диалога с пользователем в процессе вывода решений, который ведет модель на основе предложенного промпта.

Здравствуйте! Начнем диагностику системы теплоснабжения.

Для начала мне нужно уточнить некоторые параметры работы системы.

Пожалуйста, ответьте:

- Какое сейчас давление в системе теплоснабжения?

📄 👍 🗨️ ...

Нормальное

Спасибо за ответ.

Теперь уточните, пожалуйста:

- Какая сейчас температура подачи теплоносителя?

📄 👍 🗨️ ...

высокая

Следующий вопрос:

- Какая сейчас температура обратки теплоносителя?

📄 👍 🗨️ ...

низкая

Рис. 4. – Пример диалога с БЯМ, выполняющей функции ЭС с выводом решения по базе экспертных знаний (снимок с экрана с веб-доступом к модели по адресу alice.yandex.ru)

Теперь выводим рекомендации.

На основании полученных данных:

- Давление в системе нормальное ($K(\text{Факт})=1$)
- Температура подачи высокая ($K(\text{Факт})=1$)
- Температура обратки низкая ($K(\text{Факт})=1$)

Производим расчет для гипотезы D2 (Засорен теплообменник):

1. Правило 1: Если температура подачи высокая и температура обратки низкая, то возможна плохая теплоотдача

$$K(\text{ЕСЛИ}) = \min(1,1) = 1$$

$$K(\text{ТО}) = 1 * 0.75 = 0.75$$

2. Правило 2: Если возможна плохая теплоотдача и давление в норме

$$K(\text{ЕСЛИ}) = \min(0.75,1) = 0.75$$

$$K(\text{ТО}) = 0.75 * 0.8 = 0.6$$

3. Для завершения цепочки вывода необходимо уточнить:

- Когда проводилось последнее обслуживание теплообменника?

🗨️ 👍 🗨️ ...

год назад

Производим финальный расчет для гипотезы D2:

3. Правило 3: Если возможен засор теплообменника и теплообменник не обслуживался более года

$$K(\text{ЕСЛИ}) = \min(0.6,1) = 0.6$$

$$K(\text{ТО}) = 0.6 * 0.9 = 0.54$$

Вероятность гипотезы D2 (Засорен теплообменник) = 0.54

Рекомендации:

- Поскольку теплообменник засорен, а температура подачи высокая, рекомендуется:
- Провести немедленную промывку теплообменника
- После промывки проверить параметры работы системы
- Организовать регулярное обслуживание теплообменника не реже раза в год

Дополнительно рекомендуется:

- Контролировать давление в системе во время проведения работ
- При снижении давления немедленно прекратить работы и проверить систему на возможные повреждения

🗨️ 👍 🗨️ ...

Рис. 5. – Пример диалога с БЯМ (продолжение, начало на рис. 4)

На рис. 4–5 видно, что модель YandexGPT 5 Pro успешно справляется с такими задачами, как опрос пользователя для получения нужных данных; означивание и выполнение правил продукций с многошаговым выводом решений, в том числе, с вычислением вероятностных коэффициентов.

Заключение

Результаты продемонстрировали способности предлагаемой архитектуры ЭС на основе БЯМ к выполнению функций экспертной системы. Благодаря управляющему промπτу ЭС модель справляется с задачами опроса пользователя для означивания и активации правил из базы знаний, выводит решения, которые предусмотрены этой базой знаний, в том числе на основе составных цепочек вывода. Необходимо учитывать, что результаты работы такой системы зависят от структуры и содержания промπτов. Поэтому для обеспечения стабильности и повторяемости работы БЯМ с нужными результатами в промπτ добавляется блок ограничений, а также правила ведения диалога (см. Таблицу 1). Стабильность работы языковой модели с разработанным промπτом была проверена на серии из 20 экспериментов с внешней оценкой релевантности диалога правилам вывода из БЗ.

Можно предположить, что разные языковые модели будут по-разному реагировать на подобные промπτы. Поэтому при создании ЭС важно предусмотреть этап исследований и прототипирования, когда до начала программной разработки с помощью вэб-интерфейса языковой модели можно провести проверку и настройку промπτов, а также баз знаний с целью отбора наиболее рациональных БЯМ из доступных разработчикам. После такого прототипирования программная реализация системы выполняется с подключением к БЯМ через программный интерфейс (Application Programming Interface – API).

Дальнейшие исследования в этом направлении связаны с проектированием и разработкой многоагентных БЯМ-систем, где будут выделены функциональные агенты, работающие на разных этапах функционирования ЭС. В числе прочего необходимо предусмотреть реализацию агентов, выполняющих математические вычисления для гибридных, расчетно-логических систем, а также блоков оперативной и долговременной памяти для контролируемого выполнения длинных цепочек вывода и хранения результатов.

Источник финансирования

Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации (проект FEWZ-2024-0052).

Литература

1. Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research// ACM Comput. Surv. 1, 1 (January 2025), 37 pages. arxiv.org/pdf/2501.04306v1
2. Busch, F., Hoffmann, L., Rueger, C. *et al.* Current applications and challenges in large language models for patient care: a systematic review// Commun Med 5, 26, 2025. URL: doi.org/10.1038/s43856-024-00717-2
3. Sayeed Salih, Omayma Husain, Mosab Hamdan, Samah Abdelsalam, Hashim Elshafie, Abdelwahed Motwakel, Transforming education with AI: A systematic review of ChatGPT's role in learning, academic practices, and institutional adoption // Results in Engineering, Volume 25, 2025, 103837, URL: doi.org/10.1016/j.rineng.2024.103837.
4. Liu, O., Fu, D., Yogatama, D., & Neiswanger, W. DeLLMa: Decision Making Under Uncertainty with Large Language Models // International Conference on Learning Representations (ICLR), 2025. URL:

proceedings.iclr.cc/paper_files/paper/2025/file/6cd3ac24cdb789beea9f7145670fcae-Paper-Conference.pdf

5. Abram Handler, Kai R. Larsen, Richard Hackathorn, Large language models present new questions for decision support, International Journal of Information Management, Volume 79, 2024, 102811, URL: doi.org/10.1016/j.ijinfomgt.2024.102811.

6. Евлахова М.А., Зубков А.В. Обзор применения больших языковых моделей в информационных задачах систем поддержки принятия решения на примере здравоохранения // Инженерный вестник Дона. 2025. №3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9956

7. Raza, M., Jahangir, Z., Riaz, M.B. et al. Industrial applications of large language models. Sci Rep 15, 13755 (2025). URL: doi.org/10.1038/s41598-025-98483-1

8. Xuan Liu, John Ahmet Erkoyuncu, Jerry Ying Hsi Fuh, Wen Feng Lu, Bingbing Li Knowledge extraction for additive manufacturing process via named entity recognition with LLMs, Robotics and Computer-Integrated Manufacturing, Volume 93, 2025, 102900, URL: doi.org/10.1016/j.rcim.2024.102900.

9. Вакушин А.А., Клебанов Б.И. Применение больших языковых моделей в имитационном моделировании//Инженерный вестник Дона. 2024. №2. URL: ivdon.ru/ru/magazine/archive/n2y2024/8990

10. Попов Э.В. Экспертные системы: Решение неформализованных задач в диалоге с ЭВМ. – М.: Издательство «Наука», 1987. - 288 с.

11. Джексон П. Введение в экспертные системы. – М.: Издательский дом «Вильямс», 2001. – 624с.

12. Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. Базы данных. Интеллектуальная обработка информации. – М.: Издательство Нолидж, 2001. – 496с.

13. Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., Liao, L., Li, X.: Retrieval augmented generation-driven information retrieval and question answering in construction management. *Advanced Engineering Informatics*, Volume 65, Part A, 103158 (2025). URL: doi.org/10.1016/j.aei.2025.103158
 14. Sun, Y.; Liu, X. Research and Application of a Multi-Agent-Based Intelligent Mine Gas State Decision-Making System. *Appl. Sci.* 2025, 15, 968. URL: doi.org/10.3390/app15020968
 15. Dandea, V.; Grigoras, G. Expert System Integrating Rule-Based Reasoning to Voltage Control in Photovoltaic-Systems-Rich Low Voltage Electric Distribution Networks: A Review and Results of a Case Study. *Appl. Sci.* 2023, 13, 6158. URL: doi.org/10.3390/app13106158
 16. Duer, R.; Duer, S.; Zajkowski, K.; Woźniak, M.; Bernatowicz, D.; Paś, J.; Stawowy, M.; Iqbal, A.; Harničárová, M. Wind Power Plant Expert System Diagnostic Knowledge Base Creation. *Energies* 2025, 18, 1843. URL: doi.org/10.3390/en18071843
 17. Sawicki, P.; Sawicka, H.; Karkula, M.; Zajda, K. Combined Rough Sets and Rule-Based Expert System to Support Environmentally Oriented Sandwich Pallet Loading Problem. *Energies* 2025, 18, 268. URL: doi.org/10.3390/en18020268
 18. Syahputri, I.W.; Budiardjo, E.K.; Putra, P.O.H. Unlocking the Potential of the Prompt Engineering Paradigm in Software Engineering: A Systematic Literature Review. *AI* 2025, 6, 206. URL: doi.org/10.3390/ai6090206
 19. Lin, Z. How to write effective prompts for large language models. *Nat Hum Behav* 8, 611–615 (2024). URL: doi.org/10.1038/s41562-024-01847-2
 20. Aljanabi, Mohammad; Yaseen, Mohanad Ghazi; Ali, Ahmed Hussein; and Mohammed, Mostafa Abdulghafoor. Prompt Engineering: Guiding the Way to Effective Large Language Models. // *Iraqi Journal for Computer Science and Mathematics* // 2023, Vol. 4: Iss. 4, Article 12.
-

References

1. Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR. *ACM Comput. Surv.* 1, 1 (January 2025), 37 pages. URL: arxiv.org/pdf/2501.04306v1
2. Busch, F., Hoffmann, L., Rueger, C.. *Commun Med* 5, 26, 2025. URL: doi.org/10.1038/s43856-024-00717-2
3. Sayeed Salih, Omayma Husain, Mosab Hamdan, Samah Abdelsalam, Hashim Elshafie, Abdelwahed Motwakel. Volume 25, 2025, 103837, URL: doi.org/10.1016/j.rineng.2024.103837.
4. Liu, O., Fu, D., Yogatama, D., & Neiswanger, W. DeLLMa: URL: proceedings.iclr.cc/paper_files/paper/2025/file/6cd3ac24cdb789beea9f7145670fcae-Paper-Conference.pdf
5. Abram Handler, Kai R. Larsen, Richard Hackathorn. Volume 79, 2024, 102811, URL: doi.org/10.1016/j.ijinfomgt.2024.102811.
6. Evlakhova M.A., Zubkov A.V. *Inzhenernyy vestnik Dona*. 2025. №3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9956
7. Raza, M., Jahangir, Z., Riaz, M.B. *Sci Rep* 15, 13755 (2025). URL: doi.org/10.1038/s41598-025-98483-1
8. Xuan Liu, John Ahmet Erkoyuncu, Jerry Ying Hsi Fuh, Wen Feng Lu, Bingbing Li. Volume 93, 2025, 102900. URL: doi.org/10.1016/j.rcim.2024.102900.
9. Vakushin A.A., Klebanov B.I. *Inzhenernyy vestnik Dona*. 2024. №2. URL: ivdon.ru/ru/magazine/archive/n2y2024/8990
10. Popov E.V. *Ekspertnye sistemy: Reshenie neformalizovannykh zadach v dialoge s EVM* [Expert systems: Solving informal problems in dialogue with a computer]. M.: Izdatel'stvo «Nauka», 1987. 288 p.
11. Dzhekson P. *Vvedenie v ekspertnye sistemy* [Introduction to Expert Systems]. M.: Izdatel'skiy dom «Vil'yams», 2001. 624p.

12. Korneev V.V., Gareev A.F., Vasyutin S.V., Raykh V.V. Bazy dannykh. Intellektual'naya obrabotka informatsii [Databases. Intelligent information processing]. M.: Izdatel'stvo Nolidzh, 2001. 496p.
13. Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., Liao, L., Li, X. Volume 65, Part A, 103158 (2025). URL: doi.org/10.1016/j.aei.2025.103158
14. Sun, Y.; Liu, X. Appl. Sci. 2025, 15, 968. URL: doi.org/10.3390/app15020968
15. Dandea, V.; Grigoras, G. Appl. Sci. 2023, 13, 6158. URL: doi.org/10.3390/app13106158
16. Duer, R.; Duer, S.; Zajkowski, K.; Woźniak, M.; Bernatowicz, D.; Paś, J.; Stawowy, M.; Iqbal, A.; Harničárová, M. Energies 2025, 18, 1843. URL: doi.org/10.3390/en18071843
17. Sawicki, P.; Sawicka, H.; Karkula, M.; Zajda, K. Energies 2025, 18, 268. URL: doi.org/10.3390/en18020268
18. Syahputri, I.W.; Budiardjo, E.K.; Putra, P.O.H. AI 2025, 6, 206. URL: doi.org/10.3390/ai6090206
19. Lin, Z. Nat Hum Behav 8, 611–615 (2024). URL: doi.org/10.1038/s41562-024-01847-2
20. Aljanabi, Mohammad; Yaseen, Mohanad Ghazi; Ali, Ahmed Hussein; and Mohammed, Mostafa Abdulghafoor. Iraqi Journal for Computer Science and Mathematics // 2023, Vol. 4: Iss. 4, Article 12.

Дата поступления: 13.09.2025

Дата публикации: 25.10.2025